Topical Analysis of Interactions Between News and Social Media

Ting Hua*, **Yue Ning ***, **Feng Chen†**, **Chang-Tien Lu***, **Naren Ramakrishnan***

∗ Department of Computer Science, Virginia Tech
Falls Church, VA 22043
† Department of Computer Science, University at Albany-SUNY
Albany, NY 12222

Abstract

The analysis of interactions between social media and traditional news streams is becoming increasingly relevant for a variety of applications, including: understanding the underlying factors that drive the evolution of data sources, tracking the triggers behind events, and discovering emerging trends. Researchers have explored such interactions by examining volume changes or information diffusions, however, most of them ignore the semantical and topical relationships between news and social media data. Our work is the first attempt to study how news influences social media, or inversely, based on topical knowledge. We propose a hierarchical Bayesian model that jointly models the news and social media topics and their interactions. We show that our proposed model can capture distinct topics for individual datasets as well as discover the topic influences among multiple datasets. By applying our model to large sets of news and tweets, we demonstrate its significant improvement over baseline methods and explore its power in the discovery of interesting patterns for real world cases.

Introduction

Recently, online social media such as Twitter have served as tools for organizing and tracking social events (Hua et al. 2013). Understanding the triggers and shifts in opinion-driven mass social media data can provide useful insights for various applications in academia, industry, and government (Lin et al. 2013b; Tumasjan et al. 2010). However, there remains a general lack of understanding of what causes the hot spots in social media. Typically, the reasons behind the rapid spread of information can be summarized in terms of two categories: exogenous and endogenous factors (Kwak et al. 2010; Lehmann et al. 2012; Lin et al. 2013a). Endogenous factors are the results of information diffusion inside the social network itself, namely, users obtain information primarily from their online social network. In contrast, exogenous factors mean that users get information from outside sources first, for example, traditional news media, and then bring it into their social network.

Although previous works have explored both the social media and external news data datasets, few researchers have looked at the endogenous and exogenous factors based on semantical or topical knowledge. They have either sought to identify relevant tweets based on news articles (Hu et al. 2012b; Jin et al. 2011), or simply correlated the two data sources through similar patterns in the changing data volume (Tsytsarau, Palpanas, and Castellanos 2014). In fact, even within the same data source, there could be various factors that drive the evolution of information over time (Leskovec, Backstrom, and Kleinberg 2009). Exogenous factors across multiple datasets make analyzing the evolution and relationship among multiple data streams more difficult (Lin et al. 2013a).

Monitoring social media and outside news data streams in a united frame can be a practical way of solving this problem. In this paper, we propose a novel topic model, News and Twitter Interaction Topic model (NTIT), that jointly learns social media topics and news topics and subtly capture the influences between topics. The intuition behind this approach is that before a user posts a message, he/she may be influenced either by opinions from his/her online friends or by articles from news agencies. In our new framework, a word in a tweet can be responsive to the topical influences coming either from endogenous factors (tweets) or from exogenous factors (news).

Figure 1 shows an example of our problem and goals. The example introduced here is a protest happened in Mexico (Hua et al. 2013). On January 7, local government arrested 26 dogs as suspects of a murder case. Twitter users angrily demanded the release of the animals that the hashtag "#yosoycan26" (I am dog 26) became a trending topic in the following day, which finally resulted in a real-world protest on January 12. Using the new NTIT model, we attempt to address the following questions: **1) Do Twitter and news cover the same set of topics?** As can be seen from the figure, the two datasets share some common topics (e.g., topic "dog" and topic "yosoycan26"), but may also have some distinct topics of their own (e.g., topic "call for protest" only appears in the Twitter dataset). **2) For each topic, which came first, news or tweets?** Topics may display different temporal patterns in different datasets. For example, at time $t1$ topic "yosoycan26" experienced a burst in the Twitter data first, followed by a news burst on the same topic shortly afterwards at time $t2$. **3) As time goes by, how do topics affect each other?** Intuitively, topic "yosoycan26" could

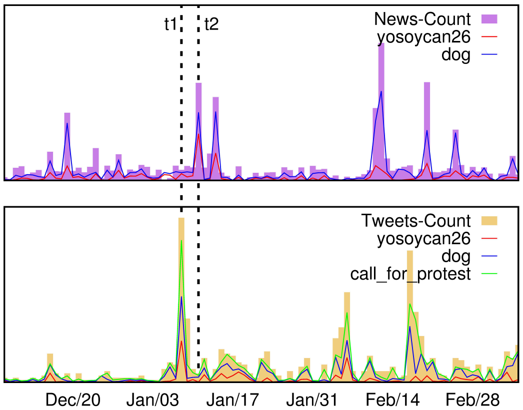


Figure 1: An example of daily volume and topics on a particular theme in News data (top) vs Tweets data (bottom). Along the timeline (x-axis), the shaded areas represent the numeric values of raw document volume for news articles and tweets; the red/blue/green curves are hidden topics discovered by our NTIT model.

be the trigger for topic "call for protest". With outputs of NTIT model, we can model such directional influence between topics quantitively. **4) What are key contributors (e.g., key documents or key player) pushing evolution of the event?** By utilizing controlling variable of NTIT, we could identify key contributors in the event evolution such as milestone news report, hot tweet, and influential users.

Our major contributions in this paper are summarized as follows:

- **We propose a novel Bayesian model that jointly learns the topics and interactions of multiple datasets.** It is already known that knowledge learned from long articles (e.g., Wikipedia) can improve the learning of topics for short messages (e.g., tweets) (Chang, Boyd-Graber, and Blei 2009; Phan et al. 2011). Our proposed model can easily transfer topical knowledge from news to tweets and improve the performance of both data sources.
- **We provide an efficient Gibbs sampling inference for the proposed NTIT model.** Gibbs sampling was chosen for the inference and parameter estimation of NTIT model for its high accuracy in estimations for LDA-like graphical model.
- **We demonstrate the effectiveness of the proposed NTIT model compared to existing state-of-the-art algorithms.** NTIT model is tested on large scale News-Twitter datasets associated with real world events. With extensively quantitative and qualitative results, NTIT shows significant improvements over baseline methods.
- **We explore real world events by using our NTIT model to reveal interesting results.** Our proposed model allows a variety of applications related to textual and temporal relationships. The learned estimations of hidden variables can be used for discoveries of various types of interests, such as topic differences, topical influences, temporal patterns, and key documents.

Table 1: Mathematical Notation

Notation	Description
$\mathbf{R}$	A set of news articles
$\mathbf{T}$	A set of tweets
$\theta_\mathbf{r}$	topic mixture proportion for news article r
$\theta_\mathbf{t}$	topic mixture proportion for tweet t
$\mathbf{Z_r}$	mixture indicator choosing topic for words in news article r
$\mathbf{Z_t}$	mixture indicator choosing topic for words in tweet t
$\mathbf{W_r}$	words in news set R
$\mathbf{W_t}$	words in tweet set T
x_t	document indicator for words in tweet to choose topics
μ_t	indicator for tweet words choosing the document to draw topics
α_r	Dirichlet parameters of Multinomial distributions θ_r
α_t	Dirichlet parameters of Multinomial distributions θ_t
α_x	Dirichlet parameters of Multinomial distributions μ_x
β	Dirichlet parameters for mixture components

Model and Inference

Beyond numeric features of raw document volume (Tsytsarau, Palpanas, and Castellanos 2014), focus of this paper is to identify underlying topics of the two data sources, and explore their relationships. Specifically, we define our problem as follows.

Problem *Given a news document set and a tweet set, tasks of this paper include: 1)measure topic coverage differences between news and Twitter; 2)reveal temporal patterns based on topics; 3)model directional influence between topics; 4) identify key contributors for the event evolution.*

Model

As shown in Figure 2, NTIT jointly models news topics and Twitter topics, under an asymmetrical frame. The generative process is described in Algorithm 1.

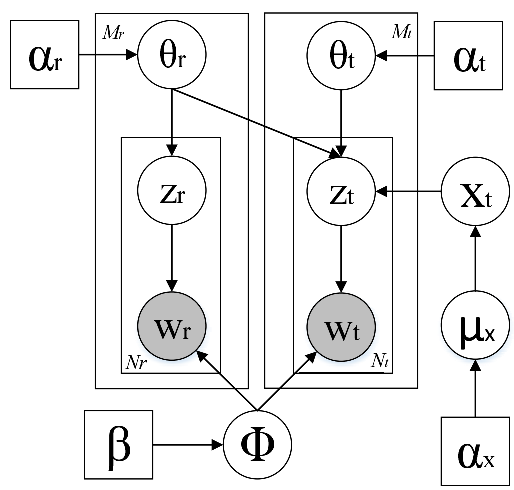


Figure 2: NTIT graphical model

An LDA-like generative process has been chosen for topic modeling in the news documents (left panel of Figure 2). To model behaviors of Twitter users (right panel of Figure 2), we assume that tweets consist of words that are either sampled from news topics (exogenous factor) or from Twitter topics (endogenous factor). Learning the hidden topics from long articles is known to be helpful for topic modeling on short texts (Phan, Nguyen, and Horiguchi 2008). However, directly applying the trained topic model derived from long texts to short messages will eliminate their distinct features, such as hashtags, mentions, and user comments. To control the influence from news to tweets, news articles and tweets are connected by a multinomial variable $\mathbf{X_t}$. If the sampled result of x_t is a document m_r from news set, the tweet word will draw its topic assignment from news document m_r. Otherwise, if x_t indicates that tweet topics have been selected, the tweet word will be generated from $Mult(\theta_t)$. The benefits of our proposed NTIT can be summarized as follows.

1. **Easy to identify common topics.** In our NTIT, a common topic term distribution ϕ is shared by both tweets and news documents, which facilitates the identification of common topics. Meanwhile, the topic variations in different datasets can be easily calculated based on their word frequency weights.
2. **Capable of retaining distinct features.** In NTIT model, tweets are able to learn enriched topics from the knowledge of long news articles while preserving their distinct features. Meanwhile, unlike symmetrical topic models (Jin et al. 2011), NTIT is an unsymmetrical model that can prevent errors and noises of tweets from impacting modeling of news documents.
3. **Useful for measuring topic influence.** Through indicator $\mathbf{X_t}$, the new NTIT model can easily tell whether a tweet word is generated from news topics or tweet topics. This new control variable can bring together topic-term distribution $\mathbf{\Phi}$ and doc-topic mixture $\mathbf{\Theta}$ and thus provide a chance to evaluate the topic level influence.

Inference via Gibbs Sampling

Gibbs sampling is chosen as the inference of the proposed NTIT model for its unbiased estimations on LDA-like graphical models (Welling and Teh 2008). Based on the generative process illustrated in Algorithm 1 and the graphical model in Figure 2, the joint distribution of NTIT model can be represented as Equation (1):

$$\begin{aligned}&P(\mathbf{Z_r},\mathbf{Z_t},\mathbf{X_t},\mathbf{W_r},\mathbf{W_t}|\alpha_r,\alpha_t,\alpha_x,\beta)\\&=\int P(\mathbf{W_r}|\mathbf{Z_r},\mathbf{\Phi})P(\mathbf{W_t}|\mathbf{Z_t},\mathbf{\Phi})P(\mathbf{\Phi}|\beta)d\mathbf{\Phi}\\&\cdot\int P(\mathbf{Z_r}|\theta_r)P(\mathbf{Z_t}|\theta_t,\mathbf{X_t}\in R)P(\theta_r|\alpha_r)d\theta_r\\&\cdot\int P(\mathbf{Z_t}|\theta_t,\mathbf{X_t}\in\mathbf{T})P(\theta_t|\alpha_t)d\theta_t\\&\cdot\int P(\mathbf{X_t}|\mu_x)P(\mu_x|\alpha_x)d\mu_x.\end{aligned}\tag{1}$$

The key to this inferential problem is to estimate posterior distributions of hidden variables $\mathbf{Z_r}$, $\mathbf{Z_t}$, and $\mathbf{X_t}$. Gibbs sampling iteratively samples one instance at a time, conditional on the values of the remaining given variables. Taking the inference of $\mathbf{Z_r}$ as an example, the Gibbs sampler estimates $P(z_{r,i}=k|\mathbf{Z_{r,\neg i}},\mathbf{Z_t},\mathbf{X_t},\mathbf{W_t},\mathbf{W_n})$ rather than the original probability, where i denotes the nth word in document m. After cancelling those factors that are independent of $z_{r,i}$, the posterior can obtained in Equation (2):

for *each topic* $k\in[1,K]$ **do**
  draw mixture component $\varphi_k\sim Dir(\beta)$;
for *each news document* $m_r\in\mathbf{M_r}$ **do**
  draw topic proportions $\theta_{m_r}\sim Dir(\alpha_r)$;
  for *each word* w_r^i *in news document* d_r **do**
    draw topic index $z_{m_r}^i\sim Mult(\theta_{m_r})$;
    draw word $w_r^i\sim Mult(\varphi_{z_{m_r}^i})$;
for *each tweet* $m_t\in\mathbf{M_t}$ **do**
  for *each word* w_t^i *in tweet* $\mathbf{m_t}$ **do**
    draw indicator $x_{t,w}\sim Mult(\mu_t)$;
    if $x_{t,w}\in\mathbf{R}$ **then**
      draw topic index $z_{m_t}^i\sim Mult(\theta_r^{(x_t)})$;
    if $x_{t,w}\in\mathbf{T}$ **then**
      draw topic proportions $\theta_t\sim Dir(\alpha_t)$;
      draw topic index $z_{m_t}^i\sim Mult(\theta_t^{(x_t)})$;
    draw word $w_t^i\sim Mult(\varphi_{z_{m_t}^i})$;

Algorithm 1: Generation Process of NTTT model

$$\begin{aligned}&P(z_{r,i}=k|\mathbf{Z_{r,\neg i}},\mathbf{Z_t},\mathbf{X_t},\mathbf{W_t},\mathbf{W_n})\\&=\frac{n_{r,w}^k+n_{t,w}^k+\beta_w-1}{\sum_{w=1}^{V}n_{r,w}^k+n_{t,w}^k+\beta_w-1}\cdot\frac{n_{r,m_r}^k+n_{t,m_r}^k+\alpha_{r,k}-1}{\sum_{k=1}^{K}n_{r,m_r}^k+n_{t,m_r}^k+\alpha_{r,k}-1},\end{aligned}\tag{2}$$

where V is the vocabulary size, $n_{t,w}^k$ and $n_{r,w}^k$ are the numbers of times that topic k is assigned to word w in tweets and news. n_{r,m_r}^k denotes the number of times topic k is assigned to words in news document m_r. And n_{t,m_r}^k is the number of times topic k appears in words of tweets, which are generated by topic mixture proportion θ_r of document m_r.

The inference of $\mathbf{Z_t}$ is slightly different from that of $\mathbf{Z_r}$, since words in tweets can be drawn from either a news document m_r or a tweet message m_t. Therefore, the conditional probability of $P(z_{t,i}=k|\mathbf{Z_r},\mathbf{Z_{t,\neg i}},\mathbf{X_t},\mathbf{W_t},\mathbf{W_n})$ can be calculated through two cases determined by the topic choosing indicator $\mathbf{X_t}$. When $\mathbf{X_t}\in\mathbf{R}$, word topic assignment $\mathbf{Z_t}$ is drawn from a multinomial distribution by the Dirichlet prior θ_r from a news document m_r:

$$\begin{aligned}&P(z_{t,i}=k|\mathbf{Z_r},\mathbf{Z_{t,\neg i}},\mathbf{X_t},\mathbf{W_t},\mathbf{W_n})\\&=\frac{n_{r,w}^k+n_{t,w}^k+\beta_w-1}{\sum_{w=1}^{V}n_{r,w}^k+n_{t,w}^k+\beta_w-1}\cdot\frac{n_{r,m_r}^k+n_{t,m_r}^k+\alpha_{r,k}-1}{\sum_{k=1}^{K}n_{r,m_r}^k+n_{t,m_r}^k+\alpha_{r,k}-1}.\end{aligned}\tag{3}$$

When $\mathbf{X_t}\in\mathbf{T}$, word topic assignment $\mathbf{Z_t}$ is drawn from a multinomial distribution by the Dirichlet prior θ_t from a tweet document m_t:

$$P(z_{t,i}=k|\mathbf{Z_r},\mathbf{Z}_{\mathbf{t},\neg\mathbf{i}},\mathbf{X_t},\mathbf{W_t},\mathbf{W_n}) = \frac{n^k_{r,w}+n^k_{t,w}+\beta_w-1}{\sum_{w=1}^{V} n^k_{r,w}+n^k_{t,w}+\beta_w-1}\cdot\frac{n^k_{t,m_t}+\alpha_{t,k}-1}{\sum_{k=1}^{K} n^k_{t,m_t}+\alpha_{t,k}-1}, \quad (4)$$

where n^k_{t,m_t} is the number of times topic k appears in the words of tweets, which are generated by topic mixture proportion θ_t of tweet m_t.

As can be seen from Algorithm 1, $\mathbf{X_t}$ is a control variable that determines whether a tweet word is sampled from a tweet message m_t or a news document m_r. To facilitate the inference, the Dirichlet distribution is chosen as the conjugate prior for $\mathbf{X_t}$. As for $\mathbf{Z_t}$, the posterior of $\mathbf{X_t}$ is discussed for two cases here. When $\mathbf{X_t}\in\mathbf{R}$, we have:

$$P(x_{t,i}=u|\mathbf{Z_r},\mathbf{Z_t},\mathbf{X}_{\mathbf{t},\neg\mathbf{i}},\mathbf{W_t},\mathbf{W_n}) = \frac{n^k_{r,m_r}+n^k_{t,m_r}+\alpha_{r,k}-1}{\sum_{k=1}^{K} n^k_{r,m_r}+n^k_{t,m_r}+\alpha_{r,k}-1}\cdot\frac{n^u_{x_t\in R}+\alpha_{x,u}-1}{\sum_{u=1}^{M_r+M_t} n^u_{x_t\in R}+\alpha_{x,u}-1}, \quad (5)$$

where $n^u_{x_t\in R}$ is the number of tweet words choosing topic mixture proportion of news document u_r. For words with $\mathbf{X_t}\in\mathbf{T}$, tweet messages are chosen as the topic mixture proportions:

$$P(x_{t,i}=u|\mathbf{Z_r},\mathbf{Z_t},\mathbf{X}_{\mathbf{t},\neg\mathbf{i}},\mathbf{W_t},\mathbf{W_n}) = \frac{n^k_{t,m_t}+\alpha_{t,k}-1}{\sum_{k=1}^{K} n^k_{t,m_t}+\alpha_{t,k}-1}\cdot\frac{n^u_{x_t\in T}+\alpha_{x,u}-1}{\sum_{u=1}^{M_r+M_t} n^u_{x_t\in T}+\alpha_{x,u}-1}, \quad (6)$$

where $n^u_{x_t\in T}$ denotes the number of tweet document u chosen as the topic mixture proportion for tweet words.

Finally, multinomial parameters $\mathbf{\Phi}=\{\varphi_k\}_{k=1}^K$, $\mathbf{\Theta_r}=\{\theta_{r,m}\}_{m=1}^{M_r}$, $\mathbf{\Theta_t}=\{\theta_{t,m}\}_{m=1}^{M_t}$, and $\mu_{\mathbf{x}}=\{\mu_u\}_{u=1}^{M_r+M_t}$ can be estimated through above posteriors, according to Bayes' rule and the definition of Dirichlet prior.

Discovery for topic lags and influence

The output results of NTIT model can be used for further discoveries, such as topic distribution differences, topic temporal patterns, topic influence, and key news documents or tweets.

Topic distribution differences

The difference of topic distribution between the two datasets can be evaluated through integrating their respective word distribution. Taking news data for instance, topic-term weight $\mathbf{D_r}$ of news documents is K-dimensional vector, where each element $D_{r,k}$ can be calculated as follows:

$$D_{r,k}=\sum_{m_r=1}^{M_r}\sum_{w_r=1}^{N_{m_r}}\varphi_{k,w}\cdot n^k_{m_r,w_r}, \quad (7)$$

where m_r denotes a specific news document, M_r is the total number of news documents, $\phi_{k,w}$ is the probability of word w in topic k, and $n^k_{m_r,w_r}$ is the count of a specific word w_r in news document m_r. Twitter topic-term vector $\mathbf{D_t}$ can be calculated in a similar way.

Topic temporal patterns

To evaluate the temporal patterns of topics, we construct topic-term time series by splitting topic-term weights (Equation (7)) through daily sliding window. Also taking news data for example, $\mathbf{T_{r,k}}=\{D_{r,k}(\tau):\tau\in\mathbf{T}\}$ represents the topic-term time series, where each element is a topic-term weight at time τ denoted by $D_{r,k}(\tau)$:

$$D_{r,k}(\tau)=\sum_{t_{m_r}=\tau}\sum_{w_r=1}^{N_{m_r}}\varphi_{k,w}\cdot n^k_{m_r,w_r}. \quad (8)$$

Instead of integrating all news documents as shown in Equation (7), $D_{r,k}(\tau)$ only considers news documents with timestamp t_{m_r} equal to τ. Twitter topic-term time series $\mathbf{T_{t,k}}$ can be calculated in a similar way.

Topic influence

In NTIT model, topics are multinomial distributions over words $\phi_{k,w}$, the topic-document indicator $\chi_{w,u}$ denotes number of times that document u is chosen by word w for generation, and $\theta_{u,k}$ implies the probability of topic k appearing in document u. By integrating the three variables and marginalizing $\phi_{k,w}$ over words, the probability of topic k_j being influenced by topic k_i can be evaluated as follows:

$$p(k_i\rightarrow k_j)=\sum_{w\in T,u\in D_R\cup D_T}\varphi_{k_i,w}\cdot\chi_{w,u}\cdot\theta_{u,k_j}. \quad (9)$$

Equation (9) provides a method to quantify the directional topic influence between any two topics, from which we can easily explain whether a topic k_j is evolved from topic k_i.

Key news reports and tweets

The topic-document indicator $u=x_{t,w}$ represents that: document u is chosen as topic mixture prior to generate tweet word w. The importance of document u can therefore be calculated by the number of words that select u as their topic mixture:

$$\mathbf{I_u}=\sum_{w=1}^{w\in\mathbf{T}}x^u_{t,w}. \quad (10)$$

The more important a document u is, the more words will refer it as their topic mixture, which can therefore yield a larger $\mathbf{I_u}$. The top ranked news reports and tweets by $\mathbf{I_u}$ are treated as key documents that dominate topics.

Experiment

In this section, we first describe our evaluation datasets, and then compare our proposed NTIT model with existing state-of-the-art algorithms. Finally, extensive discovery results are presented by exploring the outputs of NTIT.

Dataset

To construct News dataset and Twitter dataset for evaluation, we crawled publicly accessible data using RSS API and Twitter API [1]. For *news dataset*, we focus on influential civil events in Latin America. Events in this domain

[1] https://dev.twitter.com/rest/public

are chosen due to their great social influence and high evolution complexity. An event is considered "influential" if it is reported by all the top local news outlets. News reports corresponding to the event are downloaded as data for the News dataset. The tweets used for the experiments in this paper are collected via the following steps: 1) Select keywords from the title and abstract of news reports by TF-IDF; 2) retrieve relevant tweets by keywords identified in Step 1 and manually check their relevance to the given news; 3) in truly relevant tweets, identify those hashtags specifically correlated to the given news; 4) retrieve Twitter data again through the hashtags identified in Step 3. The tweets of step 2 and 4 are kept in our Twitter dataset. In total, we selected 74 influential events in the period from January 2013 to December 2013 that occurred in 5 countries in Latin America, including 1,266,548 tweets and 132,756 news reports. There are an average of 25.2 words per tweet message and 304.7 words per news article.

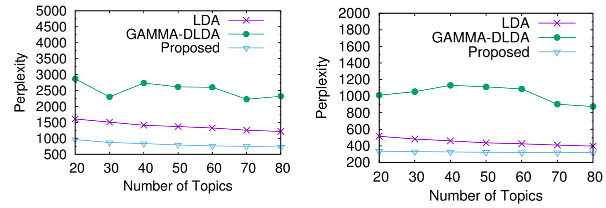
Results of modeling performance

Perplexity In this paper, we have chosen standard LDA (Blei, Ng, and Jordan 2003), Gamma-DLDA (Jin et al. 2011), and ET-LDA (Hu et al. 2012b) as baselines for comparison. These models are similar in time complexity since they are inferred through Gibbs sampling. Perplexity is therefore chosen to evaluate models' capabilities of estimating data, which typically defined as follows:

$$Perplexity(D) = exp\{\frac{-\sum_{d=1}^{M} logP(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\},$$

where M is the number of documents, $\mathbf{w}_d$ is the word vector for document d and N_d is the number of words in d. A lower perplexity indicates more accurate performance of the model.

Figure 3 presents the perplexity comparison for the 3 models on both the news and Twitter datasets. Gamma-DLDA returns high perplexity values, LDA and ET-LDA achieves intermediate performance, and our model exhibits lowest perplexity on both news and tweets. The poor performance of Gamma-DLDA is due to its completely symmetrical structure. Long articles are known to be helpful for improving the modelling performance of short messages (Phan, Nguyen, and Horiguchi 2008), but a symmetrical structure will propagate errors and noises from short texts to long texts. Unlike Gamma-DLDA, our NTIT model is unsymmetrical in structure, which can improve Twitter modelling performance through knowledge learned from news, as well as suppressing the negative impact from Twitter to news. ET-LDA is also an unsymmetrical model and therefore gets the second best performance on tweets. However, tweet words in ET-LDA can only be generated from news topics or background topics, excluding key tweet topics which are considered in NTIT. LDA is a traditional model for topic analysis, but achieve non-trivial performance on both news and Twitter. Next, we will evaluate NTIT model against the baseline method LDA in terms of semantical meaning.



(a) Perplexity of News (b) Perplexity of Tweets

Figure 3: Perplexity Comparison for News and Tweets Datasets

Semantics Table 2 presents the top words of 3 selected topics discovered from the theme "teacher protests" in Mexico. For better interpretation, the listed topics are manually assigned with meaningful labels. "CNTE", "SNTE", and "CETEG" are three teacher organizations in Mexico, which emerged as important topics in different periods throughout the series of protests. As can be seen from Table 2, NTIT and LDA display similar performance on news datasets, but yield quite different results for the Twitter dataset. This leads to several interesting observations:

1. Each topic from NTIT can be easily correlated to the corresponding label, since the representative hashtag is ranked highly. For example, in the topic "CETEG" for the NTIT model, the hashtag "#CETEG" is the top ranked word in the word list and most remaining words are directly related to the label. In contrast, it is to hard distinguish topics in LDA: (i) both "#snte" and "#cnte" appear in the topic "CNTE"; (ii) topics share too many common words, such as "government", "reform", which indicates that the LDA model tends to output unclear topic mixtures.
2. Most words identified as NTIT topics are related to the label, such as "march" and "teacher". But LDA seems to produce more meaningless background words, such as "television", "#Mexico", and "#photo".
3. Tweet topics from NTIT retain more distinct Twitter features than LDA. In addition to the key word "#ceteg", the NTIT "CETEG" topic contains event specific hashtags such as "#FebreroMesDeLaCruzada". Similar examples can also be found in the other two topics "CNTE" and "SNTE". This result demonstrates that the NTIT model is able to prevent short texts from being "submerged" by long text topics.

Results of topic evolution discovery

Topic distributions and influence Do news outlets and Twitter cover the same topics? To explore this question, we can calculate topic-term distributions using Equations (7), and the normalized results are shown in first two columns of Table 3. The results clearly show that topics distribute quite differently in Twitter and news. Topics 5, 7, 11, 12, and 19 are tweet-dominant topics that mainly appear in Twitter data (red rows), while topics 2, 3, 9, 10, 13, 16, 17 and 18 are news-dominant topics that are more likely to exist in news data (green rows); the remaining topics are common topics that are almost evenly distributed across Twitter and the news (yellow rows).

Table 2: Top words of top topics of NTIT and LDA. Words are translated from Spanish to English by Google translator.

Model&DataSet	Topics	Words
NTIT on News	CETEG	ceteg, government, aztec, freeway, access, game
	CNTE	cnte, teacher, veracruz, oaxaca, reform, coordinator, march
	SNTE	snte, teacher, national, worker, union, federal, pressure
NTIT on Tweets	CETEG	#ceteg, duty, government, education, class, fight, #FebreroMesDeLaCruzada
	CNTE	#cnte, maestro, fortnight, march, rob, #oaxaca, gabinocue, lana
	SNTE	reform, #snte, elba, educate, #educacion, arrest, national, government, duty
LDA on News	CETEG	find, want, arrive, duty, time, president
	CNTE	cnte, teacher, reform, government, city, education, national
	SNTE	snte, drink, find, agreement, class
LDA on Tweets	CETEG	#ceteg, televise, government, #mexico, reform, #photo, support, ask, education
	CNTE	#cnte, teacher, #snte, reform, march, education, government, law
	SNTE	teacher, #snte, reform, #photo, education, government, national, ask

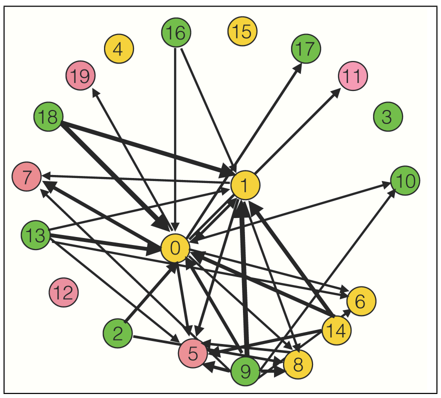


Figure 4: Topic Influence. As in Table 3, yellow nodes are common topics, red nodes are tweet-dominant topics, and green nodes are news-dominant topics. The directions of the arrows imply the directions of influence.

To further explore the relationships between topics, we can apply Equation (9) to calculate the topic influence, producing the results shown in Figure 4. Each node in Figure 4 represents a topic and the correlations between topics are denoted by the width of the edges. Edges with widths below a certain threshold (e.g., 0.15) are ignored. Node features such as degrees, in-degree ratio, and out-degree ratio are listed in the last three columns of Table 3. Compared with the news-dominant topics and Twitter-dominant topics, common topics are more likely to have greater numbers of connections, such as topic 0 (No.1 in degree) and topic 1 (No.2 in degree). News-dominant topics have a strong influence on other topics, with 62% edges being outgoing arrows. In contrast, tweet-dominant topics are weak in influence that none have an outgoing edge. These observations mirror the real world situation: news agencies can easily lead public opinion, while the voice of individuals is almost negligible.

Temporal Patterns Many researchers believe that Twitter data are disseminated earlier than traditional media when spreading news (Hu et al. 2012a). Is this true? To answer this question, we can quantitatively compare the temporal difference between the Twitter and news topics. Time series are first calculated through Equation (8), after which peaks can be detected using pypeaks [2].

Table 3: Topic Influence. "Twitter %" is the ratio of topic in Twitter data, while "News%" is the ratio of topic in news data. "Degree" denotes the node degree for each topic,"In%" is the ratio of incoming edges, and "Out%" is proportion of out-going edges.

Topic	Twitter %	News	Degree	Out%	In%
0	0.56	0.44	16	0.6	0.4
1	0.44	0.56	10	0.4	0.6
2	0.34	0.66	2	1	0
3	0.26	0.84	0	0	1
4	0.48	0.62	0	0	1
5	0.92	0.08	6	0	1
6	0.48	0.52	3	0	1
7	0.91	0.09	3	0	1
8	0.47	0.53	5	0.2	0.8
9	0.35	0.65	7	1	0
10	0.36	0.64	2	0	1
11	0.89	0.11	1	0	1
12	0.67	0.30	0	0	1
13	0.32	0.68	4	1	0
14	0.45	0.55	3	1	0
15	0.49	0.51	0	0	1
16	0.26	0.74	2	1	0
17	0.27	0.73	1	0	1
18	0.32	0.68	2	1	0
19	0.92	0.08	1	0	1

Results for the topic temporal features are listed in Table 4. Looking at the last row of Table 4, the Twitter data come slightly earlier than News in terms of bursts, with an average lead time of 0.36 hours. Red rows denote the topics that appeared earlier in tweets, with larger values in positive peak ratio. Green rows are topics that showed up first in the news, with higher negative ratios. Yellow rows indicate topics with approximately simultaneous peaks.

Interesting patterns can be obtained by correlating Table 3 with Table 4. Generally, 5 out of 7 common topics in Table 3 are also simultaneous topics in Table 4, 4 out of 5 topics that

[2] https://github.com/gopalkoduri/pypeaks

Table 5: Top 5 key news documents in “teacher protests” theme. Texts are translated from Spanish to English by Google translator.

News ID	importance count	news report title
985	478	CNTE prepare to build the united organization
4243	414	Politics at play in Mexico’s ongoing teacher protests
1684	409	Teachers’ movement: faces and reasons for fighting
5453	351	SNTE creative protest against the constitutional reform
8468	347	Protesters in 14 states join the protest CNTE

Table 6: Top 5 key tweets in “teacher protests” theme. Texts are translated from Spanish to English by Google translator.

Tweet ID	importance	content	author
413114	91	Bullying also occurs from student to teacher: SNTE leader	proceso
332824	41	teachers retired in protest because they pay them their retirement insurance, accuses indifference SECC 28 of SNTE	SoledadDurazo
38974	32	#EnVivo The eviction of members of #CNTE http://bit.ly/1aI8AeQ #Eventos #news #Nacional #DF #Maestros #Protesta	AgendaFFR
136883	17	CNTE members marched on Reforma and Bucareli to Segob where assembled tents to install a sit	REFORMACOM
39368	15	The # socket and was evicted by police, congratulations can now celebrate their “independence” and “freedom”. #CNTE	josemiguelgon

Table 4: Comparison of topic temporal patterns. “Pos%” denotes the ratio of peaks occurring earlier in Twitter than in news, “Neg%” implies that peaks appeared earlier in the news, and “Sim%” indicates the ratio of peaks that burst simultaneously in the two datasets. “Avg.Lag” indicates the average time lags between news and Twitter peaks, where positive values imply Twitter data come first while negative numbers denote the leading time of news data.

Topic	Pos%	Neg%	Sim %	Avg. Lag
0	0.20	0.30	0.50	-0.60
1	0.36	0.27	0.36	-1.09
2	0.21	0.50	0.29	-1.14
3	0.25	0.33	0.42	-1.33
4	0.40	0.20	0.40	-0.20
5	0.00	0.00	0.00	0.53
6	0.44	0.22	0.33	0.30
7	0.44	0.56	0.00	-0.25
8	0.47	0.18	0.35	2.12
9	0.33	0.25	0.42	1.00
10	0.43	0.14	0.43	0.22
11	0.36	0.29	0.36	-0.57
12	0.22	0.44	0.33	-2.67
13	0.44	0.44	0.11	0.22
14	0.40	0.30	0.30	0.40
15	0.58	0.33	0.08	2.00
16	0.54	0.15	0.31	3.69
17	0.18	0.36	0.45	-1.82
18	0.00	0.60	0.40	-2.40
19	0.41	0.18	0.41	2.82
Total	0.35	0.31	0.34	0.36

first in the news in Table 4 are news-dominant topics in Table 3, and 4 out of 5 topics that show up first in the news in Table 4 are either tweet-dominant topics or common topics in Table 3. Outliers are topic 12 and topic 16, which are in fact essential for the understanding of interaction between news and Twitter. Topic 12 is a Twitter-dominant topic in Table 4, which would thus be expected to appear first in tweets but in fact occurs earlier in the news data. Top ranked words in topic 12 include: “educate”, “elba” (name of the leader of SNTE), and “arrest”. By manually checking corresponding news and tweets, we found that: at the end of 2013 February, the leader of SNTE “Elba Esther Gordillo” was arrested by the Mexican government because of corruption allegations. This event was just a regular news report for news agencies, but unexpectedly attracted great attentions from social media users, and actually became the main trigger of many of the subsequent protests. Topic 16 is a news-dominant topic in Table 4, that shows up first in tweets. Top ranked words in topic 16 include: “march”, “oaxaca”, and temporal terms such as “12:30pm”. Obviously, items in topic 16 can be regarded as organized events that developed from virtual social media first and then caught the attention of traditional media once events began to occur in the real world.

Key news reports and tweets Table 5 and Table 6 present the top ranked key news articles and tweets respectively, according to the importance calculated using Equation (10).

News documents are more frequently cited by words than tweets. As can be seen from the Table 5 and Table 6, news documents have hundreds of references, while even most popular tweet messages are only cited less than 20 times. This is quite reasonable since news documents are much longer and have more words than tweet posts. It is also clear that the key news articles listed are representative, largely because they are either the most updated movement reports (e.g., News 985) or for the comprehensive event analysis they provide (e.g., News 1684).

Interesting results can be found in the key tweets listed in Table 6. Most of these top ranked tweets are posted by key players, such as celebrities or authoritative media. For example, tweet 332824 is posted by a user named "Soledad Durazo", a famous journalist in Mexico. Other key tweets contain numerous keywords, such as tweet 38974, which basically consists of a set of popular hashtags.

Conclusion

In this paper, we have proposed a hierarchical Bayesian model NTIT to analyze the topical relationship between news and social media. Our model enables jointly topic modeling on multiple data sources in an asymmetrical frame, which benefits the modeling performance for both long and short texts. We present the results of applying NTIT model to two large-scale datasets and show its effectiveness over non-trivial baselines. Based on the outputs of NTIT model, further efforts are made to understand the complex interaction between news and social media data. Through extensive experiments, we find following factors: 1) even for the same events, focuses of news and Twitter topics could be greatly different; 2) topic usually occurs first in its dominant data source, but occasionally topic first appearing in one data source could be a dominant topic in another dataset; 3) generally, news topics are much more influential than Twitter topics.

Acknowledgement

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract D12PC00337. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US government.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. In *The Journal of Machine Learning Research*, volume 3, 993–1022. MIT Press.

Chang, J.; Boyd-Graber, J.; and Blei, D. M. 2009. Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 169–178. ACM.

Hu, M.; Liu, S.; Wei, F.; Wu, Y.; Stasko, J.; and Ma, K.-L. 2012a. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2751–2754. ACM.

Hu, Y.; John, A.; Wang, F.; and Kambhampati, S. 2012b. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 59–65. AAAI.

Hua, T.; Lu, C.-T.; Ramakrishnan, N.; Chen, F.; Arredondo, J.; Mares, D.; and Summers, K. 2013. Analyzing civil unrest through social media. volume 46, 80–84. IEEE.

Jin, O.; Liu, N. N.; Zhao, K.; Yu, Y.; and Yang, Q. 2011. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 775–784. ACM.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, 591–600. ACM.

Lehmann, J.; Gonçalves, B.; Ramasco, J. J.; and Cattuto, C. 2012. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, 251–260. ACM.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Memetracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 497–506. ACM.

Lin, S.; Wang, F.; Hu, Q.; and Yu, P. S. 2013a. Extracting social events for learning better information diffusion models. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 365–373. ACM.

Lin, Y.-R.; Margolin, D.; Keegan, B.; and Lazer, D. 2013b. Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd International Conference on World Wide Web*, 737–748. ACM.

Phan, X.-H.; Nguyen, C.-T.; Le, D.-T.; Nguyen, L.-M.; Horiguchi, S.; and Ha, Q.-T. 2011. A hidden topic-based framework toward building applications with short web documents. volume 23, 961–976. IEEE.

Phan, X.-H.; Nguyen, L.-M.; and Horiguchi, S. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, 91–100. ACM.

Tsytsarau, M.; Palpanas, T.; and Castellanos, M. 2014. Dynamics of news events and social media reaction. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 901–910. ACM.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fifth International AAAI Conference on Web and Social Media*, 178–185. AAAI.

Welling, M., and Teh, Y. W. 2008. Hybrid variational/gibbs collapsed inference in topic models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 587–594. AUAI.