Discovering Newsworthy Tweets with a Geographical Topic Model

[Extended Abstract]

James McInerney[a,c], David M. Blei[a,b,c]
[a]Computer Science, [b]Statistics, [c]Institute for Data Sciences and Engineering
Columbia University, New York
james@cs.columbia.edu, david.blei@columbia.edu

ABSTRACT

Newsworthy events are regularly reported on Twitter in real time by eyewitnesses. However, identifying and summarising large numbers of tweets to assist journalists in discovering newsworthy information is an open problem. In this paper, we propose a probabilistic model and inference scheme that identifies the topical, geographical, and temporal structure of *latent events* that explain large numbers of observed tweets. We explore the potential for discovering newsworthy tweets using this method, with preliminary experiments involving messages collected from Upper Manhattan, New York during 24 hours. Specifically, we find that we can recover the latent group corresponding to a real newsworthy event, the Harlem, New York explosion of March 2014, from tweets obtained as the event unfolded.

1. INTRODUCTION

Newsworthy events are regularly reported on Twitter in real time by eyewitnesses. For example, a helicopter crashed into a crane in London one morning in January 2013 and was tweeted about almost immediately [1]. In March 2014, an explosion and resulting building collapse in Harlem, New York was reported by eyewitness on Twitter minutes after it happened [2]. In addition, journalists discover crucial information about longer running events from social media, such as civil unrest (e.g., the UK riots in 2011) and even revolution (e.g., the Egyptian revolution in 2011) [8]. The potential for discovering news from Twitter is clear. However, identifying and summarising newsworthy tweets is a challenging problem because the majority of messages on Twitter are not newsworthy, and reports that might be deemed newsworthy cannot always be trusted.

To overcome this, a crucial signal that a significant event has indeed happened is the occurrence of tweets that share similar geographical positions, timestamps, and word content. The frequency of tweets exhibiting such similarities provides a measure of importance (i.e., important events tend to be reported independently by lots of people) and cross-verification (i.e., it is harder for multiple users to falsify events compared to a single user). Clearly, this type of signal may be detected using clustering, and this is indeed the most popular method for event detection using Twitter data [9, 7].

However, we identify two shortcomings of clustering as the sole solution for identifying newsworthy tweets. Firstly, the range of vocabulary to describe the same event is varied (e.g., "explosion" v.s. "loud noise", or "police" v.s. "emergency services"), meaning that tweets with very similar semantic content might not be detected as such by clustering. Similarly, events are often multi-faceted, in the sense that they consist of sub-events (each with its own set of eyewitnesses) that together make up a breaking news story. For example, early emergency responders arrive at a location, injured people are found and treated, traffic is halted in the surrounding neighbourhood. In light of this, applying a mixed-membership model such as latent Dirichlet allocation (LDA) [4] to tweets is an obvious extension, but does little to address this issue in practice. This is because LDA relies on co-occurrence of words within documents, but tweets are always short (limited to 140 characters), limiting the effectiveness of LDA. Secondly, and more fundamentally, a data journalist would surely value a ranking of detected events by *newsworthiness*, which is determined by more than frequency of messages alone (e.g., dozens of people might tweet about a private function they are attending without the concert being deemed newsworthy by outsiders).

These two considerations (the need to identify semantically similar tweets and provide a measure of newsworthiness) motivate the use of an external data source during analysis. For example, if the New York Times is "all the news that's fit to print" then the corpus of New York Times articles clearly has some bearing on the semantic similarity and newsworthiness of reports on Twitter. In this paper, we propose a method of transfer learning that discovers hidden structure in one dataset (i.e., the New York Times corpus) and uses that structure to focus inference and learning with tweets.

Our contributions are the following:

- We propose a new model for automatically identifying and characterising latent newsworthy events on Twitter.
- We provide a practical inference schedule for learning about topics from an existing news corpus and then transferring this knowledge to discover geographical, temporal, and topical similarities between tweets.
- Applying our model to a small set of approximately

NewsKDD '14 New York, NY, USA

2,000 real tweets collected over a 24 hour period that includes reports of the Harlem explosion and building collapse, the newsworthy tweets are identified as a group after running inference on the model.

- We find that the topic distribution of the inferred newsworthy event is significantly closer to the topic distribution of the actual news article describing the event than any of the other inferred events in the dataset. This provides preliminary evidence that sets of tweets may be correctly matched to news articles by topic similarity.
- We identify future directions for improving research into geographic-topic analysis of newsworthy tweets.

The rest of this paper is structured as follows. We outline our proposed model and methodology in Section 2. Applying our method to a small Twitter dataset, we find preliminary evidence of the effectiveness of our model in Section 3. Finally, we draw conclusions and discuss ways to fix the shortcomings of our findings in Section 4.

2. LATENT VARIABLE MODEL OF NEWSWORTHY TWEETS

In this section we develop our model of newsworthy tweets. We start by describing our latent event model in Section 2.1. In Section 2.2, we extend this model to perform transfer learning between news corpora and tweets, in order to identify newsworthy tweets. We finish this section with details on model inference in Section 2.3.

2.1 Latent Event Model for Tweets

In this subsection, we present our generative model of tweets that extends the standard clustering assumption for tweets. We assume there exist a set of K discrete *latent events* $\boldsymbol{e}$ that are responsible for generating the observed locations $\boldsymbol{l}$, words $\boldsymbol{w}$, and timestamps $\boldsymbol{t}$ of all tweets:

$$
\begin{aligned}
&\boldsymbol{\pi} \sim DP(\alpha_\pi) \text{ draw latent event frequency coefficients}\\
&\text{for each event } k \in [1..K]:\\
&\quad \boldsymbol{\mu}_k \sim NIW(\boldsymbol{\alpha_\mu}) \text{ draw location parameters for event } k\\
&\quad \tau_k \sim \mathcal{N}(\boldsymbol{\alpha_\tau}) \text{ draw the mean timestamp for event } k\\
&\quad \boldsymbol{\psi}_k \sim DP(\boldsymbol{\alpha_\theta}) \text{ draw topic proportions for event } k\\
&\text{for each tweet } n \in [1..N]:\\
&\quad e_n \sim \mathcal{M}(\boldsymbol{\pi}) \text{ draw the event assignment for tweet } n\\
&\quad l_n \sim \mathcal{N}(\boldsymbol{\mu}_{e_n}) \text{ draw the location for tweet } n\\
&\quad t_n \sim \mathcal{N}(\tau_{e_n}) \text{ draw the timestamp for tweet } n\\
&\quad \text{for each word } m \in [1..M]:\\
&\quad\quad z_{n,m} \sim \mathcal{M}(\boldsymbol{\psi}_{e_n}) \text{ draw } m^{\text{th}} \text{ topic for tweet } n\\
&\quad\quad w_{n,m} \sim \mathcal{M}(\boldsymbol{\beta}_{z_{e_{n,m}}}) \text{ draw } m^{\text{th}} \text{ word for tweet } n
\end{aligned}
\tag{1}
$$

where bold symbols represent a matrix of values, and α_π, $\boldsymbol{\alpha_\theta}$, $\boldsymbol{\alpha_\mu}$, $\boldsymbol{\alpha_\tau}$ are fixed hyperparameters. Locations $\boldsymbol{l}$ are the observed latitude and longitude of tweets, words $\boldsymbol{w}$ are word token identifiers, and timestamps $\boldsymbol{t}$ are in seconds since the epoch (January 1, 1970 at midnight UTC).

The model given in Equation 1 is amenable to standard Bayesian inference methods, e.g., Gibbs sampling, variational inference [3]. In particular, distributions over all but one of the unknown random variables may be approximated given the observed locations, words, and timestamps of the tweets. However, we require the topic-word distribution $\boldsymbol{\beta}$. To obtain this, we perform transfer learning from an existing news corpus to this tweet model.

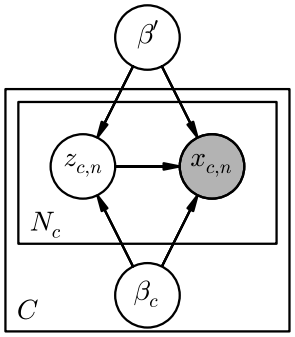

Figure 1: Plate diagram of the hierarchical latent variable model for transfer learning. Each node represents a random variable, arrows represent conditional dependencies, and shaded nodes represent observed values.

2.2 Transfer Learning from News to Tweets

We now briefly define a general approach to transfer learning for latent variable models, (a large class of probabilistic models that includes mixture models, LDA, hidden Markov models, Kalman filters [3]) that is applicable to our problem of newsworthy tweet discovery using an existing news corpus. The three sets of random variables in a standard latent variable model are: the global parameters $\boldsymbol{\beta}$, N observations $\boldsymbol{x}_n$, and N local latent variables and parameters $\boldsymbol{z}_n$ (where $n \in [1, N]$) [5]. To induce transfer learning between different corpora[1], we split the global parameters $\boldsymbol{\beta}$ into corpus-specific parameters $\boldsymbol{\beta}_c$ (for each corpus c) and truly global parameters $\boldsymbol{\beta}'$. In addition, we rename the local observations and parameters $\boldsymbol{x}_{c,n}$ and $\boldsymbol{z}_{c,n}$, respectively (and assume N_c observations per corpus, and C corpora in total). The result is a *hierarchical latent variable model*, depicted in Figure 1. The shared global parameters $\boldsymbol{\beta}'$ act as an information channel between corpora.

The channel between corpora in our newsworthy tweet discovery problem consists of information about topics, specifically, the word distribution for each topic $\boldsymbol{\beta}'$ discovered from an existing standard news corpus (e.g., New York Times, Washington Post) using LDA (corresponding to the corpus-specific latent variable model). The word distributions $\boldsymbol{\beta}'$ are also used in the mixture model detailed in Section 2.1 (so that there are 2 corpora in the hierarchical model, i.e., $C = 2$ in Figure 1).

2.3 Inference Procedure

In this subsection we describe the inference procedure we use for the models described in Sections 2.1 and 2.2. To obtain the shared topic parameters $\boldsymbol{\beta}'$, we first run inference on an existing news corpus (1,728,000 New York Times articles for our experiments) using variational inference. Variational inference is an approximate inference method that optimises

[1]N.B., each corpus might have a different assumed model structure, but share some global random variables.

Table 1: Real tweets about the Harlem, New York explosion (on 116th Street and Park Avenue)

Location (to nearest intersection)	Time & Date	Text
117th Street and Park Avenue	9:32am 03/12/2014	Explosion on park and 116th Harlem New York
111th Street and Park Avenue	9:32am 03/12/2014	Holy s***. If you in east Harlem, did ya hear that loud as sound?
119th Street and Madison Avenue	9:33am 03/12/2014	#Explosion on 116 and park in #Harlem... what is going on? Whole building shook. This is bad. Lots of smoke.
Unknown (assigned near explosion)	9:34am 03/12/2014	Just some kind of explosion in Harlem. Looks like around 5th Ave and 110 judging by the smoke I see out my window
118th Street and Lexington Avenue	9:35am 03/12/2014	I can see people running down 117 towards Park yelling 'come on! Let's go!' Smoke is clearing.
121st Street and Park Avenue	9:35am 03/12/2014	Explosion in building near 116 and park. All at Bailey house ok. We are hoping or neighbors are ok too.
121st Street and Lexington Avenue	9:37am 03/12/2014	Huge explosion at east harlem. A building just completely exploded.
112th Street and Lexington Avenue	9:40am 03/12/2014	Building blow up in 116 street park ave smh @ NYCHA - Jonson Houses

the difference between the posterior distribution of the true model and a factorisation of the posterior [6]. Unfortunately, standard variational inference uses coordinate ascent, so is not suitable for large corpora because even a single pass over the data takes $\mathcal{O}$(hours) to complete. We therefore use stochastic variational inference (SVI) with LDA[2], using the equations and algorithm of Hoffman et al. 2013 [5]. SVI performs gradient ascent on the variational objective using randomly subsampled batches of the whole corpus.

Using the $\boldsymbol{\beta}^{'}$ learnt from this procedure (keeping it fixed), we then apply variational inference to the tweet model described in Section 2.1. The numbers of tweets we consider in this paper are of the scale of 1,000s and are amenable to coordinate ascent variational inference. This would clearly not be the case for a deployed system but we leave the task of applying SVI to the tweet model for future work.

3. EXPERIMENTAL EVALUATION

In this section, we describe our experimental procedure and results for testing the ability of our approach to identify and summarise newsworthy tweets.

3.1 Experimental Procedure

For our experiments we use a set of 2,535 geotagged tweets collected with the Twitter Streaming API[3] from the Upper Manhattan area of New York over a 24 hour period starting 7pm Saturday 21 June 2014. We convert the words of these tweets into a bag of words representation, restricted to the vocabulary already obtained and used for the New York Times corpus. Removing tweets with no words in that vocabulary yields $N = 1961$ tweets, which we converted into the form (w_n, l_n, t_n) for $n \in [1, N]$ (described in Section 2.1).

Checking the dataset manually, we find no newsworthy events. Therefore, we inserted 8 real tweets from eyewitness of the March 2014 Harlem explosion written as the event unfolded (obtained from [2]). These are shown in Table 1. To insert them into the aforementioned larger set of tweets, we process them in the same way, and also modify the date to Sunday 22 June 2014 to make the event less identifiable (otherwise, the 8 tweets would form their own temporal cluster, so we would not even need to consider word content or location). All 8 tweets each have at least 3 words from our predefined vocabulary, so none are discarded during processing.

We perform two sets of experiments to determine our algorithm's ability to identify the Harlem explosion as a newsworthy event from the tweets alone:

- **Experiment 1**: Insert the Harlem explosion tweets into the full tweet dataset and see if the algorithm can discover the event. Look at the associated topics and mean location and check with ground truth.
- **Experiment 2**: Find the difference between the topic distributions of full text news articles and those of the inferred events, using the Kullback-Leibler divergence (a non-symmetric measure of difference between probability distributions). See if the event corresponding to the Harlem explosion stands out in any way. We varied the full text news articles under consideration:
  - **Comparison 1**: A New York Times article about the explosion published 12th March 2014.
  - **Comparison 2**: 10,000 randomly selected news articles from the New York Times

  None of these comparison documents appears in the original news corpus that we used to learn the topics.

3.2 Results

In Experiment 1 (described in Section 3.1) we ran coordinate ascent variational inference to convergence, yielding 69 latent events with at least 0.5 tweets assigned to them. All 8 tweets about the Harlem explosion were assigned to a single event, event 57 (along with 10 other tweets with at least 99% probability). The mean latitude/longitude of this inferred event was (40.799418, -73.941140), which is 198 meters from the location of the site of the Harlem explosion, supporting the intuition that the locations of eyewitnesses tend to be noisy estimates of an event's true location (which was a modelling assumption we made in Section 2.1). The top 5 topics associated with event 57 are given in Table 2. The remaining 10 tweets assigned to the same event (that did not describe the Harlem explosion) were not newsworthy, but had similar locations and times to the 8 newsworthy tweets, and many of them contained references to places nearby (e.g., "nyc",

[2]N.B., we assume 300 topics.

[3]https://dev.twitter.com/docs/api/streaming

Table 2: Top 5 topics for inferred event 57

Topic strengths	Topic counts	Top 10 words
0.0171	6.602	notice deaths beloved memorial paid loving funeral devoted wife family
0.0167	6.048	fire killed port killing guard emergency jordan prince marine injuries
0.0166	6.056	manhattan east brooklyn village queens neighborhood greenwich yorkers jefferson jamaica
0.0166	6.049	kind feel pretty feeling thing comfortable wonder forget wish feels
0.0148	3.894	building apartment buildings bedroom bath market taxes rent apartments room

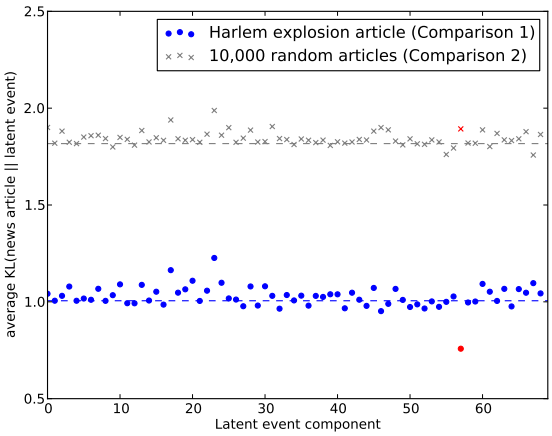

Figure 2: Topic differences between New York Times articles and the set of inferred latent events. We compare against two different sets of articles. The red markers correspond to the inferred event describing the Harlem explosion (at x-axis position 57). The horizontal dashed lines indicate the topic difference between the articles and the uniform distribution.

"east", "harlem", "lexington"), matching the third most popular topic for the event. We discuss in Section 4 how we might improve the model to avoid grouping irrelevant tweets with newsworthy tweets. However, on the whole, reducing the number of non-newsworthy tweets under consideration from almost 2,000 to 10 is a promising outcome for this experiment.

In Experiment 2, we compare the expected topic proportions (which are normalised) for all the inferred events against three sets of real news articles. Comparison 1 compares the inferred events against the New York Times write-up about the Harlem explosion (published on the same day). The result of the comparison is shown with the circular dots in Figure 2, in which we see that the inferred event corresponding to the Harlem explosion is significantly closer to the news article than any other inferred event for this dataset. For Comparison 2, we randomly select 10,000 New York Times articles and find the average similarity to their inferred topics and those of the event topics (shown by the crosses in Figure 2). This experiment has a negative result, in the sense that the inferred event reporting the Harlem explosion is not close, on average, to the random set of news articles. We hypothesise that this is because further grouping of the news articles is required (either through supervised or unsupervised means), in order to further distinguish between breaking news (e.g., involving explosions, emergency services), editorials, obituaries, blogs etc. We leave this as a focus for future work (outlined in Section 4).

Interestingly, the closest event on average to the set of articles (which can be identified from Figure 2 as event 55), contained tweets about music and partying, with the top tweet referencing Make Music New York, a music event celebrating the summer solstice held 21 June 2014[4]. We did not anticipate that this event would appear in the dataset. Make Music New York is a biannual event that is regularly covered in the arts and entertainment sections of the major New York newspapers, providing preliminary evidence of the potential for using the KL divergence as a measure of newsworthiness, given the set of inferred events and a corpus of news articles.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we identified some open problems in discovering and characterising newsworthy reports from Twitter, and proposed a model and a methodology for transferring knowledge of topics obtained from an existing news corpus to discovering newsworthy tweets. Using a small, geographically focused dataset, we showed that our method is able to identify and summarise a group of newsworthy tweets reporting an explosion in Harlem, New York, that were hidden among almost 2,000 irrelevant tweets. Results from comparing the topic distribution of this inferred event to the topic distributions of existing news articles suggest that future newsworthy events might be identified by this means.

However, a larger-scale study is clearly needed in order to confirm the validity of our findings, including a comparison with competing methods. Additionally, we identify two outstanding issues that arose during our experiments.

First, how can we further refine the discovered events so that fewer irrelevant tweets are included in them? For example, in our experiment we found that the 8 tweets about the explosion were grouped with 10 irrelevant tweets.

Second, part of the task of characterising latent events is surely to provide an explanation to the user of the type of news they describe. We found that the KL divergence from the news article describing the event was a good metric to identifying the newsworthy tweet event, but comparing to a larger set of news articles yielded a negative result. We hypothesise that the negative result occurred because the larger set of articles contains a mixture of breaking news, editorials, obituaries, blogs etc., and we would not expect an event to be close to all of them in latent space[5]. If this is correct, then the next best step is likely to be subgrouping the news corpus (by supervised or unsupervised means) before comparing topic distributions.

5. ACKNOWLEDGMENTS

Many thanks to Allison Chaney, Jeremy Manning, and Jonathan Stray for their helpful suggestions.

[4]http://makemusicny.org/about/overview/

[5]We also tried taking the min instead of the mean of the similarities, and excluding news articles without strong topic expressions, both without success.

6. REFERENCES

[1] London helicopter crash: two die in Vauxhall crane accident. *BBC News Online*, January 2013.

[2] T. Bailey. How Twitter confirmed the explosion in Harlem first. *Giga Om*, March 2014.

[3] C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

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

[5] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[6] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[7] C. Li, A. Sun, and A. Datta. Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM)*, pages 155–164. ACM, 2012.

[8] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton. Can twitter replace newswire for breaking news? In *Procedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013.

[9] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 605–613. ACM, 2013.