Applications of Topics Models to Analysis of Disaster-Related Twitter Data

Kirill Kireyev
kireyev@colorado.edu

Leysia Palen
palen@cs.colorado.edu

Kenneth M. Anderson
kena@cs.colorado.edu

Department of Computer Science
University of Colorado - Boulder

Abstract

The use of microblogging (using tools like Twitter and SMS messaging) during disasters offers a valuable source of information for disaster response agencies, as it often provides critical up-to-date and on-location updates about an unfolding crisis. This precipitates an interest in robust processing and visualization tools. We explore the use of Topics models for analysis of disaster-related Twitter data. We experiment with Topics-based clustering and visualization, corpus selection, term weighting, as well as a new technique called dynamic corpus refinement.

1. Introduction: Microblogging, Twitter and Disaster Research

Microblogging is a form of lightweight chat that allows users to send short messages to people subscribed to their streams. Microblogging services include Twitter, Jaiku, Plurk, me2DAY, among several others. Twitter, on which we focus here, allows its users to send short messages (140 characters or less) to others. These messages ("tweets") can be sent and retrieved through a variety of means and front-end clients, including text messaging, e-mail, the web, and other third-party applications, which are enabled through Twitter's public API.

Research on social media in disaster events is growing, and ranges from examination of common photo repositories ([5]) to social networking sites ([6]). More specifically, interest in microblogging in emergency management activities is on the rise (e.g. [8], [9], and [7]). Early research shows ([9]) that critical up-to-date and on-location updates can be found in microblog messages about an unfolding crisis, precipitating an interest in robust processing and visualization tools. This is the motivation for the preliminary research presented here.

1.1. NLP Challenges

Microblogged messages are short, heterogeneous and noisy. As such, Twitter data presents several challenges to traditional natural language processing technologies, such as:

- **Esoteric language and grammar.** Twitter messages often lack proper written grammar and punctuation. Tweets frequently employ a newspaper headline style that omits articles and auxiliary verbs. They frequently contain abbreviations, including Internet slang such as "omg" for "Oh my God!" Traditional NLP tools such as parsers are often inadequate for processing this type of data.
- **Message length.** The short messages contain very little lexical redundancy.
- **Locale-specific references.** Messages sometimes refer to specific location, events and other named entities, as well as implied references to locations ([9]). Thus, one cannot rely on pre-defined entity lists or complex named entity recognition methods.

In this paper we explore the use of Topics models for processing Twitter data. Topics models are probabilistic models originally developed for analyzing the semantic content of large document corpora. We suggest that the family of Topics models is a particularly promising

tool for analyzing of Twitter data, for some of the following reasons:

- **Bag-of-words.** Topics models are usually "bag-of-words" models, meaning that they do not rely on syntactic structure or word order in language (though can be adapted for doing so [2]) Thus, they are likely to be better suited to handle esoteric language and irregular grammar of typical Twitter messages.

- **Latent variables.** Topics models are able to infer latent (or hidden) relationships between elements in data. This makes them more robust to handling misspellings, acronyms, terminology and other variations in the surface form of messages. It also potentially allows them to derive interesting patterns and clusters in data along dimensions that may be different than researchers' intuitions might suggest.

- **Representation.** Topics models represent statistical knowledge as homogenous numerical vectors (e.g. multinomial probability distributions). This lends them to easy comparisons, visualization as well as mathematical manipulations, such as clustering.

- **Adaptability.** Because of their unsupervised nature Topics models can be easily retrained on a text corpus that is specifically adapted for a particular domain using widely available text collections, such as news or educational texts. This can result in a more refined model, a possibility we discuss in this work.

2. Data

We collected two datasets from the Twitter microblogging service on two natural disaster events that overlapped in time, both occurring on September 30, 2009. The first event was a magnitude 8.0 earthquake that occurred near the islands of American Samoa, triggering three subsequent tsunami waves that together struck the entire cluster of islands. The second event was a magnitude 7.6 earthquake that struck the city of Padang on Indonesia's island of Sumatra. Both events resulted in human casualties and massive damage to the built environment. We collected data using the Twitter Search API. We refer to the two datasets as Tsunami (T) and Earthquake (E). Their sizes are 19,829 and 23,354 messages, respectively. Average message length was 17 words (standard deviation ~6.6).

3. Supplementary Corpus Selection

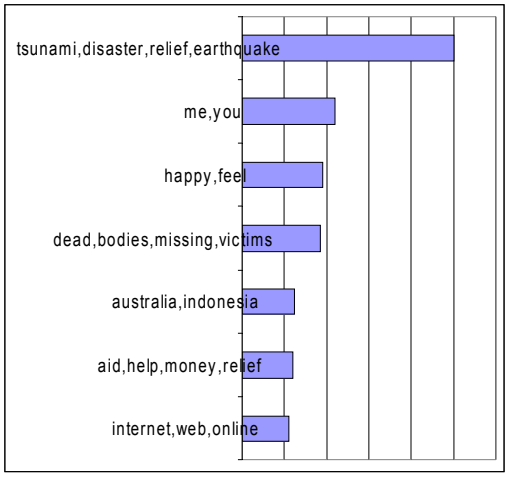
Typically, Topics models are trained and tested on data of similar nature. However, individual Twitter messages are quite short (17 words on average), and do not contain enough statistical redundancy for inferring word relationships. Alternatively, we experiment with training the Topics models on general-purpose corpora (we experiment with a ~44,000 document TASA corpus ([4]) and a comparably-sized random subset of the ACQUAINT-2 [10] news corpus), and subsequently applying them to Twitter messages. We find that such corpora are appropriate for high-level visualization and clustering, as seen in the next section. We propose a technique for automatic corpus refinement in Section 7.

4. Qualitative Analysis

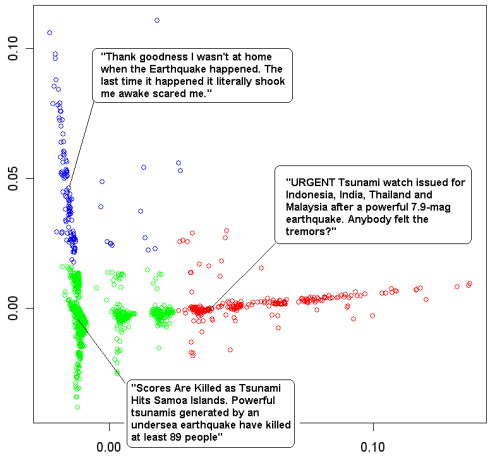
As a first step, we apply the basic LDA Topics model to the Twitter data to see if high-level patterns of salient topics correspond to our intuition as well as past qualitative research observations. The model used in these examples is a 300-topic LDA ([1]) model trained on a news corpus ([10]) with inference using Gibbs sampling. Topics of Twitter messages were computed by "folding-in:" including them into the inference, but keeping global word probabilities unchanged during iterations over these messages.

To illustrate, the Figure 1A shows the most prominent topics in the collection of Earthquake (E) twitter messages (ordered by topic prominence), together with most salient words to illustrate each topic. The salient topics correspond to intuition and show possibilities for using Topics models for visualization of Twitter messages.

To extend the illustration, the Figure 1B shows results of multi-dimensional scaling (2-dim.) and k-means clustering (3 clusters) of data, together with representative messages for each cluster. The MDS and clustering were performed using dot products between topic distributions for each message as the distance metric. In the graph below one can visually identify two main latent patterns in the messages: informational (e.g. "Urgent: Tsunami watch issued…") and emotional ("Thank goodness I wasn't at home…").



1A. Prominent topics in E-dataset



1B. MDS and clustering of E-dataset

5. Improving Term Weighting

Because Twitter messages are short, low-content words and stopwords comprise a significant part of each message and significantly influence their representation in the Topics model. To give content words more weight, we change the term weighting scheme used to compute a document's topic distribution (Θ_d). The original LDA model ([1]) implicitly uses a uniform scheme, where each word is given an equal weight. Instead, we use a term weighting scheme which is based on word's *specificity*: more specific words have a higher weight when Θ_d is computed from individual topic assignments in the document (Twitter message). We use the scheme described in [3], which is based on word vector length in Latent Semantic Analysis:

$$spec(w) = log(||v_w|| / f_w) + c$$

where v_w is the LSA word vector for word w, f_w is its frequency in the training corpus and c is a constant needed to make the values positive.

Furthermore, because of significant relative variations of message size in words, the contribution of content words on Θ_d may become diluted in longer messages. To correct for that, we do not normalize the topic vector, i.e.:

$\theta_j^d = \frac{n_{z(w)=j}^{(d)} + \alpha}{n_{\cdot}^{(d)} + T\alpha}$ (original Θ_d)	$\theta^{*d}_j = \sum_{z(w)=j} spec(w) + \alpha$ (modified Θ_d)

These term weighting modifications distort some basic assumptions about multinomial probability distributions implicit in LDA, but they nevertheless work well for clustering and visualization of our data.

6. Dynamic Corpus Refinement

The data for a particular disaster event contains domain- and event- specific terms such as (1) locations, (2) organization names, (3) Internet slang and (4) abbreviations. These terms may change significantly between particular disaster events. Thus, a single general-purpose training corpus may not be appropriate for every analysis. For example, the general Topics model used in Section 4 is not refined enough to assign the words "tsunami" and

"earthquake" to significantly different topic distributions.

Instead, we explore a way to dynamically refine the training corpus in an automated and unsupervised manner, so that it contains more content relevant to the particular disaster event. Consequently, the corresponding Topics model will be better able to represent the details of the Twitter messages we are interested in. To accomplish this, we:

1. Use the model built on a generic corpus as a bootstrap, compute the topic distributions on a Twitter message collection specific to a particular event.
2. Use these topic distributions to select a subset of documents from a larger database (we used ACQUAINT-2 news corpus [10]) whose topic distributions resemble the (10) most salient topics in the event-specific Twitter messages.
3. Replace a randomly-selected portion (~25%) of the default training corpus in Step 1 with selection in Step 2. Re-compute the Topics model on the training corpus, and use it to compute Topics representation of Twitter messages.

By comparing the refined model to the original one, we can observe some encouraging differences. For example, the generic model built on a news corpus generally assigns the words "tsunami" and "earthquake" to the same topic ($\phi_{tsunami} = \phi_{earthquake}$), whereas the refined model assigns them to different topics. In addition some relevant geographic locations previously unknown to the original model (such as "Padang"), become meaningfully incorporated as a result of this corpus refinement.

7. Evaluation

As with most unsupervised methods, quantitative evaluation is challenging since the objective function is not clear. Our two (T and E) datasets provide a good evaluation basis, since they correspond to two distinct disasters of similar nature (earthquake vs. tsunami) and nearby regions (Indonesia vs. Samoa). We can evaluate the impact of our variations in the Topics models by measuring how well the models discriminate between T- and E-messages. Another method is to find messages with keywords that are known to correspond to particular categories (e.g. emotional words: "god", "please" or particular locations, e.g. "Padang") and measure how well these messages are separated from the rest using pairwise comparisons or clustering. Finally, a significant part of our evaluation is qualitative analysis.

8. References

[1] D. Blei, A. Ng, and M. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[2] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum (2005). Integrating topics and syntax. In *Neural Information Processing Systems*, 2005.

[3] K. Kireyev (2009). Semantic-based Estimation of Term Informativeness. *NAACL 2009.*

[4] T. K. Landauer, S. T. Dumais (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review, 104,* 211–240.

[5] S. Liu, L .Palen, J. Sutton, A. Hughes, S. Vieweg (2008). In Search of the Bigger Picture: The Emergent Role of On-Line Photo-Sharing in Times of Disaster. In Proceedings of *ISCRAM 2008*.

[6] L. Palen, S. Vieweg (2008). Emergent, Widescale Online Interaction in Unexpected Events: Assistance, Alliance and Retreat. In Proceedings of *CSCW 2008*.

[7] L. Palen, S. Vieweg, S. Liu, A. Hughes (to appear 2009). Crisis in a Networked World: Features of Computer-Mediated Communication in the April 16, 2007 Virginia Tech Event. Social Science Computing Review, Sage, (Pages TBA).

[8] K. Starbird, L. Palen, A. Hughes and S. Vieweg (to appear 2010). Chatter on The Red:What Hazards Threat Reveals about the Social Life of Microblogged Information. ACM CSCW 2010.

[9] S. Vieweg, A. Hughes, K. Starbird and L. Palen (submitted). Supporting Situational Awareness in Emergencies Using Microblogged Information. *ACM Conf. on Human Factors in Computing Systems* 2010.

[10] ACQUAINT-2 News Corpus, Linguistic Data Consortium, *http://www.ldc.upenn.edu/*