Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

Characterizing Microblogs with Topic Models

Daniel Ramage
Stanford University
353 Serra Mall, Stanford, CA
dramage@cs.stanford.edu

Susan Dumais
Microsoft Research
One Microsoft Way, Redmond, WA
sdumais@microsoft.com

Dan Liebling
Microsoft Research
One Microsoft Way, Redmond, WA
danl@microsoft.com

Abstract

As microblogging grows in popularity, services like Twitter are coming to support information gathering needs above and beyond their traditional roles as social networks. But most users' interaction with Twitter is still primarily focused on their social graphs, forcing the often inappropriate conflation of "people I follow" with "stuff I want to read." We characterize some information needs that the current Twitter interface fails to support, and argue for better representations of content for solving these challenges. We present a scalable implementation of a partially supervised learning model (Labeled LDA) that maps the content of the Twitter feed into dimensions. These dimensions correspond roughly to substance, style, status, and social characteristics of posts. We characterize users and tweets using this model, and present results on two information consumption oriented tasks.

Introduction

Millions of people turn to microblogging services like Twitter to gather real-time news or opinion about people, things, or events of interest. Such services are used for social networking, e.g., to stay in touch with friends and colleagues. In addition, microblogging sites are used as publishing platforms to create and consume content from sets of users with overlapping and disparate interests. Consider a hypothetical user *@jane* who follows user *@frank* because of the latter's posts about college football. However, *@frank* additionally uses Twitter to coordinate social arrangements with friends and occasionally posts political viewpoints. Currently, *@jane* has few tools to filter non-football content from *@frank.* In short, Twitter assumes that *all* posts from the people *@jane* follows are posts she wants to read. Similarly, *@jane* has a limited set of options for identifying new people to follow. She can look at lists of users in the social graph (e.g. those followed by @frank), or she can search by keyword and then browse the returned tweets' posters. However, it remains difficult to find people who are like *@frank* in general or – more challengingly – like *@frank* but with less social chatter or different political views.

The example above illustrates two of the many content-oriented information needs that are currently unmet on Twitter. Solving these challenges will require going beyond the traditional network-based analysis techniques that are often applied to microblogs and social networks to develop new tools for analyzing and understanding Twitter content. Content analysis on Twitter poses unique challenges: posts are short (140 characters or less) with language unlike the standard written English on which many supervised models in machine learning and NLP are trained and evaluated. Effectively modeling content on Twitter requires techniques that can readily adapt to the data at hand and require little supervision.

Our approach borrows the machinery of latent variable topic models like the popular unsupervised model Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). Latent variable topic models have been applied widely to problems in text modeling, and require no manually constructed training data. These models distill collections of text documents (here, tweets) into distributions of words that tend to co-occur in similar documents – these sets of related words are referred to as "topics." While LDA and related models have a long history of application to news articles and academic abstracts, one open question is if they will work on documents as short as Twitter posts and with text that varies greatly from the traditionally studied collections – here we find that the answer is yes. In this paper, we use Labeled LDA (Ramage, et al., 2009), which extends LDA by incorporating supervision in the form of implied tweet-level labels where available, enabling explicit models of text content associated with hashtags, replies, emoticons, and the like.

What types of patterns can latent variable topic models discover from tweets? The *Understanding Following Behavior* section argues that the latent topics can be roughly categorized into four types: *substance* topics about events and ideas, *social* topics recognizing language used toward a social end, *status* topics denoting personal updates, and *style* topics that embody broader trends in language usage. Next, in the *Modeling Posts with Labeled LDA* section**,** we outline some applications of the mixture of latent and labeled topics, demonstrating the specificity of learned vocabularies associated with the various label types. Then, in the *Characterizing Content on Twitter* section, we characterize selected Twitter users along these learned dimensions, showing that topic models can provide interpretable summaries or characterizations of users' tweet streams. Finally, In the *Ranking Experiments* section, we demonstrate the approach's effectiveness at modeling Twitter content with a set of experiments on users' quality rankings of their own subscribed feeds.

Related work

Most of the published research about Twitter has focused on questions related to Twitter's network and community structure. For example, (Krishnamurthy, Gill, & Arlitt, 2008) summarize general features of the Twitter social network such as topological and geographical properties, patterns of growth, and user behaviors. Others such as (Java, et al., 2007), argue from a network perspective that user activities on Twitter can be thought of as information seeking, information sharing, or as a social activity.

Less work has presented a systematic analysis of the textual content of posts on Twitter. Recent work has examined content with respect to specific Twitter conventions: @user mentions in (Honeycutt & Herring, 2009) and re-tweeting, or re-posting someone else's post in (boyd, Golder, & Lotan, 2010). Notably, (Naaman, Boase, & Lai, 2010) characterizes content on Twitter and other "Social Awareness Streams" via a manual coding of tweets into categories of varying specificity, from "Information Sharing" to "Self Promotion." Naaman, et al., extrapolate from these categories, inducing two kinds of users: "informers" that pass on non-personal information and "meformers" that mostly tweet about themselves. Others have proposed forms of content analysis on Twitter with specific focuses, such as modeling conversations (Ritter, Cherry, & Dolan, 2010). Although rich with insight, these works do not present automatic methods for organizing and categorizing all Twitter posts by content, the problem we approach here.

Understanding Following Behavior

What needs drive following and reading behavior on Twitter, and to what extent does Twitter satisfy them? To help organize our own intuitions, we conducted in-depth structured interviews with four active Twitter users (with number of following and followed users ranging from dozens to thousands), and followed up with a web-based survey of 56 more users. We found that both the content of posts and social factors played important roles when our interviewees decided whether to follow a user. Distilling our conversations down to their essence, we found that all those interviewed made distinctions between people worth following for the subjects they write about (*substance,* e.g. about a hobby or professional interest), because of some social value (*social*, e.g. for making plans with friends), because of (dis)interest in personal life updates from the poster (*status*, e.g. where someone is or what they are doing), or because of the tone or style of the posts (*style*, e.g. humor or wit).

To examine these intuitions in a broader context, we conducted a web-based survey cataloging reasons that underlie users' following decisions on Twitter, as determined from our interviews and other direct interaction with regular Twitter users. 56 respondents within Microsoft completed the survey during one week in November 2009. 65% were male and 75% were between the ages of 26 and 45. 67% were very active consumers of information, reading posts several times a day. 37% posted more than once per day, and 54% posted with frequency between once a day and once a month. While this sample does not represent the full range of Twitter's demographics, we believe it provides useful insight into challenges facing Twitter users more generally.

Respondents were asked how often they considered 26 reasons when making decisions about whom to follow, with most reasons falling into one of the substance, status, social and style categories identified earlier. Each respondent rated each reason on a five-point scale: "rarely," "sometimes", "about half the time," "often," to "almost always." The most common reasons for following represent a mixture of the four categories of reasons: the two most common reasons were "professional interest" and "technology" (*substance*). These particular substantive topics reflected the demographics of the respondents. The next most commonly used reasons were "tone of presentation" (*style*), "keeping up with friends" (*social*), "networking" (*social*), and "interested in personal updates" (*status*). Low ranked reasons included "being polite by following back" and "short-term needs (like travel info)."

Respondents were also queried about nine reasons for un-following users, i.e. removing users from their streams. We found that "too many posts in general" was the most common reason for a user to be un-followed. Other common reasons were: "too much status/personal info" (*status*), "too much content outside my interest set" (*substance*), and "didn't like tone or style" (*style*). Respondents rarely un-followed for *social* reasons like "too many conversations with other people." The least common reason was, unsurprisingly, "not enough posts" – because such users are rarely seen by their followers simply by lack of activity. 24 users provided additional reasons for un-following: 10 mentioned spam, 8 mentioned insufficiently interesting / boring / duplicative posts, and 6 un-followed because of offensive posts (e.g. religious or political views, general tone, or about other people).

In response to an open-ended question about what an ideal interface to Twitter would do differently, survey respondents identified two main challenges related to content on Twitter, underscoring the importance of improved models of Twitter content. First, new users have difficulty discovering feeds worth subscribing to. Later, they have too much content in their feeds, and lose the most interesting/relevant posts in a stream of thousands of posts of lesser utility. Of the 45 respondents who answered this question, 16 wanted improved capabilities for *filtering* of their feeds by user, topic on context (e.g., *"organize into topics of interest", "ignore temporarily people, tags or topics"*). In addition, 11 wanted improved interfaces for *following*, such as organization into topics or suggestions of new users to follow (e.g. *"suggestions on who to follow that have similar interests"*).

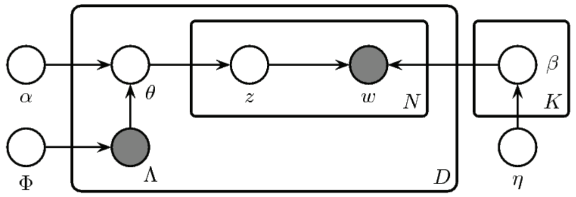
Modeling Posts with Labeled LDA

The information needs outlined above point to the importance of developing better models of textual content on Twitter. The approach we use here is based on latent variable topic models like LDA (Blei, et al., 2003). LDA is an unsupervised model that discovers latent structure in a collection of documents by representing each document as a mixture of latent topics, where a topic is itself represented as a distribution of words that tend to co-occur. LDA can be used to discover trends in language usage (what words end up together in topics) as well as to represent documents in a low-dimensional topic space.

We use a generalization of LDA, Labeled LDA (Ramage, et al., 2009), which extends LDA by incorporating supervision where available. Labeled LDA assumes the existence of a set of labels Λ, each characterized by a multinomial distribution β_k for $k \in 1..|\Lambda|$ over all words in the vocabulary. The model assumes that each document d uses only a subset of those labels, denoted $\Lambda_d \subseteq \Lambda$, and that document d prefers some labels to others as represented by a multinomial distribution θ_d over Λ_d. Each word w in document d is picked from a word distribution associated one of that document's labels, i.e. from β_z for some $z \in \Lambda_d$. The word is picked in proportion both to how much the enclosing document prefers the label $\theta_{d,z}$ and to how much that label prefers the word $\beta_{z,w}$. In this way, Labeled LDA can be used for *credit attribution* – it can attribute each word in a document to a weighted mix of the document's labels, with other words in the document helping to disambiguate between label choices.

Figure 1 shows the Bayesian graphical model and generative process for Labeled LDA. From this generative process assumption, an approximate inference algorithm can be used to reconstruct the per-document distributions θ over labels and the per-label distributions β over words, starting from only the documents themselves. Implementation details are described later in this section.

Labeled LDA allows us to model a collection of Twitter posts as a mixture of some *labeled* dimensions as well as the traditional *latent* ones like those discovered by LDA. Although not discussed in (Ramage, et al, 2009), LDA is a special case of Labeled LDA. We can model K latent topics as labels named "Topic 1" through "Topic K" assigned to every post in the collection. If no other labels are used, this label assignment strategy makes Labeled LDA mathematically identical to traditional LDA with K topics. However, Labeled LDA gives us the freedom to introduce labels that apply to only some subsets of posts, so that the model can learn sets of words that go with particular labels, like hashtags, which we will return to in the *Labeled Dimensions in Twitter* subsection.



For each topic k in 1..K, draw a multinomial distribution β_k from symmetric Dirichlet prior η.
For each tweet d in 1..D:

1. Build a label set Λ_d describing the tweet from the deterministic prior Φ
2. Select a multinomial distribution θ_d over the labels Λ_d from symmetric Dirichlet prior α.
3. For each word position i 1..N in tweet d
   - **a.** Draw a label $z_{d,i}$ from label multinomial θ_d
   - **b.** Draw a word $w_{d,i}$ from word multinomial β_z

Figure 1: Bayesian graphical model of Labeled LDA (top), and description of the model's generative process (bottom).

Dataset Description

We trained models on data collected by crawling one week of public posts from Twitter's "spritzer" stream. This public stream's makeup is determined by Twitter and contains posts sampled from all public posts made on the site. Our collection contains 8,214,019 posts from the 17th through the 24th of November 2009 (*OneWeek*). Posts were processed by tokenizing on whitespace and on punctuation subject to rules designed to keep together URLs, emoticons, usernames, and hashtags. Some multi-word entity names were collapsed into single tokens (such as michael_jackson) by using a gloss lookup derived from Wikipedia and query logs. After processing, posts contained an average of 13.1 words from a vocabulary of 5,119,312 words. As an important pre-processing step, we removed the 40 most common terms in the corpus[1] and all terms appearing in fewer than 30 documents. Some experiments were conducted on just those posts from the 24th of November (*OneDay*), containing just over 1M posts. It is worth noting that the number of documents in both collections is substantially larger than most applications of latent variable topic models, where collections tend to be on the order of tens of thousands of documents, although those documents are usually longer.

Besides the number and types of labels used, Labeled LDA has two parameters: we used un-tuned symmetric Dirichlet priors of .01 for η and .01 for α, which can be thought of as pseudo-count smoothing on per-label word distributions and per-post label distributions, respectively. In early experimentation with these values, we found similar qualitative results across a wide range of small positive values.

[1] The most common terms are effectively a corpus-specific collection of stop-words; removing them improves running time and the subjective quality of learned topics.

Model Implementation and Scalability

In order to scale to our test collection size – and beyond for real-time analysis of all Twitter data – our implementation of Labeled LDA must be parallelizable. Unfortunately, the Gibbs sampling algorithm for Labeled LDA proposed in (Ramage, et al. 2009) is inherently sequential. While parallelized approximations to Gibbs sampling for regular LDA have been proposed (Asuncion, Smyth, & Welling, 2008), we developed a simpler alternative inference algorithm based on a variational approximation to the Labeled LDA objective. Our algorithm is modeled on the CVB0 variational approximation to LDA described in (Asuncion, et al. 2009). For each word at position *i* in each post *d*, the algorithm stores a distribution $\gamma_{d,i}$ over the likelihood that each topic generated that word in that document. These distributions are then converted into counts of how often each word is paired with each label globally, denoted $\#_{kw}$, and how often each label appears in an each document, denoted $\#_{dk}$. The algorithm alternates between assigning values to $\gamma_{d,i,k}$ and then summing assignments in a counts phase. The update equations are listed below. Initially, we use small random values to initialize $\#_{kw}$ and $\#_{dk}$. The references to $\gamma_{d,i,k}$ on the right side of the proportionality in the assignment phase refer to the value at the previous iteration.

Assign: $$\gamma_{d,i,k} \propto \frac{\#_{kw}\quad \gamma_{dik}+\eta}{\#_{k}\quad \gamma_{dik}+W\eta} \cdot (\#_{dk} - \gamma_{dik} + \alpha) \cdot I[k \in \Lambda_{\mathrm{d}}]$$

Count: $$\begin{aligned} \#_{dk} &= \textstyle\sum_i \gamma_{d,i,k} \\ \#_{kw} &= \textstyle\sum_{d,i} \gamma_{d,i,k} \cdot I\left[w_{d,i} = w\right] \\ \#_{k} &= \textstyle\sum_w \#_{kw} \end{aligned}$$

Formulating our Labeled LDA learning problem in this way allows for a data-parallel implementation. Documents are distributed across a cluster of compute nodes. Before each assignment phase, all nodes are given a copy of the current counts $\#_{dk}$, $\#_{kw}$ and $\#_{k}$. The assignments phase is done in parallel on all processors. Then, processors aggregate their local counts by summing their assignments in parallel, and then passing along the sums to higher rank nodes until the master node has the sum of all counts. This iterative process repeats for a fixed number of iterations or until the change in model parameters falls below a threshold. Our implementation does threading within compute nodes and communicates across nodes with MPI, and can complete training on the OneWeek dataset within about four days on a 24-machine cluster.

In the results presented in this paper, the Labeled LDA models will contain 100 or 200 dimensions (a parameter we set) that correspond to latent trends in the data ("Topic 1" through "Topic K" applied to each post), and about 500 *labeled* dimensions (depending on the dataset) that correspond to hashtags, etc, as described in the *Labeled Dimensions in Twitter* subsection. After describing the characteristics of these dimensions, we go on to describe how they can be used to characterize users or sets of posts (*Characterizing Content on Twitter*) and how they impact performance on two ranking tasks (*Ranking Experiments*).

Latent Dimensions in Twitter

Before examining the types of content captured by the labels in Labeled LDA, we first examine Twitter's latent structure, as modeled using *K* labels applied to every post in the collection. These labels are incorporated so that unsupervised large-scale trends can be captured by the model. By inspection, we find that many of these learned latent dimensions can be divided into one of the four categories defined above: those about events, ideas, things, or people (*substance*), those related to some socially communicative end (*social*), those related to personal updates (*status*), and those indicative of broader trends of language use (*style*). Later, we refer to text analyses using these categories as a *4S analysis*.

We manually labeled 200 latent dimensions from one run of our model on the *OneDay* dataset according to the 4S categories by examining the most frequent words in each dimension's term distribution. Four raters labeled each dimension as any combination of *substance, status, style, social*, or *other* – i.e. each dimension may have more than one 4S category assignment. As an example, the most frequent words in "Topic 1" are: "watching tv show watch channel youtube episode and season," which was labeled as *substance*. The *other* dimensions tended to be dominated by non-English terms, by numbers, by symbols, or by generic word classes like terms for males (him his he boy father man, etc).

Table 1 summarizes the number of latent dimensions associated with each category, the inter-rater agreement in labeling, and the top words in an example dimension for each category. We used Fleiss' κ to compute inter-rater agreement for each of these categories across our four judges as separate binary classification tasks. As shown in Table 1, we find fair to substantial agreement across all categories. The social category shows the lowest inter-rater agreement, which is in part because so much language usage on Twitter has some social component, regardless of whether it is also substantive, stylistic, etc. Indeed, (boyd, Golder, & Lotan, 2010) report that 36% of posts mention another user, and of those roughly 86% are directed specifically to that user. As a caveat, categorizing latent dimensions in this way can be difficult for three reasons. First, the judgments (and even our categories) are inherently subjective, although we do find reasonable agreement. Second, some legitimate trends may be hidden in the lower frequency terms in each distribution. Finally, many discovered dimensions are inherently ambiguous in usage, such as some indicative linguistic styles being coupled with social intent. Nonetheless, we believe that this type of high-level summary can provide value insofar as it quantifies agreed-upon intuitions, and holds up to scrutiny when examined at the level of individual posts. In our own exploration, we found the 4S categorization corresponded to distinctions that arose commonly in the interviews, survey and content analysis and, furthermore, that there was good agreement about categorization decisions from multiple labelers.

Table 1: Inter-rater agreement from four raters marking 200 latent dimensions with 4S categories. Left: number of dimensions in category marked by >=2 raters. Middle: Fleiss' κ showing all four categories have at least fair agreement. Right: high scoring words in an example from each category.

Category	Fleiss' κ	Example topic
Substance 54/200	.754	obama president american america says country russia pope island failed honduras talks national george us usa
Status 30/200	.599	am still doing sleep so going tired bed awake supposed hell asleep early sleeping sleepy wondering ugh
Style 69/200	.570	haha lol :) funny :p omg hahaha yeah too yes thats ha wow cool lmao though kinda hilarious totally
Social 21/200	.370	can make help if someone tell me them anyone use makes any sense trying explain without smile laugh
Other 47/200	.833	la el en y del los con las se por para un al es una su mais este nuevo hoy

Labeled Dimensions in Twitter

While the latent dimensions in Twitter can help us quantify broad trends, much additional meta-data is available on every post that can help uncover specific, smaller trends. In addition to the latent dimensions discussed above, several classes of tweet-specific labels were applied to subsets of the posts. For instance, we create one label for each *hashtag*. A hashtag is a Twitter convention used to simplify search, indexing, and trend discovery. Users include specially designed terms that start with # into the body of each post. For example a post about a job listing might contain the term #jobs. By treating each hashtag as a label applied only to the posts that contain it, Labeled LDA discovers which words are best associated with each hashtag. Common words better described by some latent dimension tend not to be attributed to the hashtag label.

We incorporated several other types of labels into the model. *Emoticon-specific* labels were applied to posts that used any of a set of nine canonical emoticons: smile, frown, wink, big grin, tongue, heart, surprise, awkward, and confused. Canonical variations were collapsed: e.g.] and :-) mapped to :). *@user* labels were applied to posts that addressed any user as the first word in the post, as per the Twitter convention of direct messaging. *reply* labels were added to any post that the Twitter API has designated as a reply, i.e. because a user clicked a reply link on another post. *question* labels were applied to posts that contain a question mark character. Because the emoticons, @user, reply, and question labels were relatively common, each of these labels was factored into 10 variants – e.g. ":)-0" through ":)-9" – in order to model natural variation in how each label was used. The number 10 was chosen heuristically given the relative commonality of these symbols compared to hashtags. Posts contained an average of 8.8 labels out of a label vocabulary of 158,223 distinct labels. Of those labels, the majority (158,103) were hashtags; we filtered hashtags occurring on less than 30 posts, resulting in a final set of 504 labels.

Table 2: Example word distributions learned for various classes of labels, supplementing latent topics (not shown)

Emoticons	:)	thanks thank much too hi following love very you're welcome guys awww appreciated ah
		love all guys tweet awesome x nice twitter your goodnight followers later y'all sweet xoxo
	:(	miss sick still feeling ill can't much today already sleep triste him baby her sooo fml
		ah working won't stupid why anymore :'(isn't suck computer isnt ahh yeah nope nothing
Social Signal	Reply	thanks i'm sure ok good will i'll try yeah cool x fine yes definitely hun yep glad xx okay
		lmao yea tho yu wat kno thats nah hell lmfao idk dont doin aint naw already ima gotta we
	@user	haha yeah that's know too oh thats cool its hahaha one funny nice though he pretty yes
	?	did how does anyone know ?? ?! get where ??? really any mean long are ever see
		?! ?? !? who wtf !! huh ??? hahaha wow ?!! ?!? right okay ??!! hahahaha eh oh knew
Hashtags	#travel	travel #traveltuesday #lp hotel #ac ac tip tips #food national air airline #deals countries #tips
	#twilight	#newmoon #twilight twilight watching edward original watch soundtrack Jacob tom cruise
	#politics	#cnn al gore hoax climategate fraud #postrank gop inspires policy because why new bill

Table 2 shows some characteristic topics associated with each label class. Natural variation in the linguistic usage is evident: one of the excerpted smile labels is used to express gratitude and another consists of various forms of social bonding ("xoxo" means hugs and kisses). Similarly, one frown label is dedicated to feeling ill, whereas another represents frustration (mostly with computers). The specificity of these labeled dimensions hints at new directions in sentiment analysis on Twitter content. One *reply* label is dedicated to confirmations (thanks ok good yeah) and another represents a somewhat rowdier linguistic style (lmao yea tho wat hell). Analogous distinctions are found through the other label types. We are interested in exploring applications of isolating each of these trends, such as improved browsing interfaces for hashtag labels, better sentiment analysis using emoticon labels, and conversation and question modeling using the social labels. An open challenge in formulating this kind of model is how best to select the number of sub-labels per label type, which we plan to explore in future work.

Beyond the inherent appeal of explicitly modeling these label types, their incorporation supports our 4S analysis. For example, we know that all posts that are replies or are directed to specific users are, to some extent, *social*, so we can count usage of any *reply* or *@user* label as usage of the *social* category. Emoticons are usually indicative of a particular *style* and/or a *social* intent. Because hashtags are intended to be indexed and re-found, they might naturally

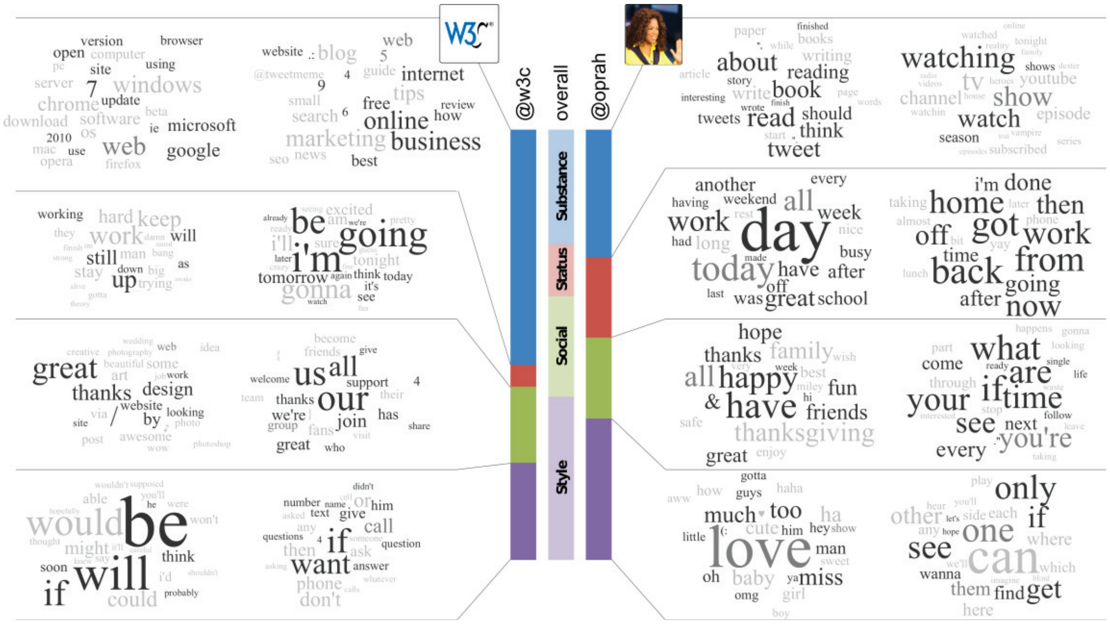


Figure 2: 4S analysis of two users: @w3c (left) and @oprah (right). The usage of dimensions from substance (top row), status (second), social (third), or style (bottom) categories is shown in the vertical bars, with Twitter's average usage shown in the center. Common words in selected dimensions from each category are shown as word clouds. Word size is proportional to frequency in that dimension globally, and word shade is proportional to the frequency in the user's recent tweets. Light gray words are unused in recent tweets.

be labeled as *substance*. Although not all labels fall cleanly into the assigned categories, the great majority of usage of each label type is appropriately categorized as listed above, enabling us to expand our 4S label space without manual annotation.

Characterizing Content on Twitter

Labeled LDA can be used to map individual posts into learned latent and labeled dimensions, which we have grouped into 4S categories – *substance status style social*, either manually (for 200 latent dimensions) or by construction (for 504 labeled ones). These mappings can be aggregated across posts to characterize large-scale trends in Twitter as well as patterns of individual usage. Formally, a post d's usage of topic k, denoted $\theta_{d,k}$ is computed simply as $\#_{dk}$ / $|d|$. We compute an aggregate signature for any collection of posts by summing and normalizing $\#_{dk}$ across a collection of documents, such as posts written by a user, followed by a user, the result set of a query, etc. The usage of any 4S category can be determined by summing across dimensions within that category.

By aggregating across the whole dataset, we can present a large-scale view of what people post on Twitter. At the word level, Twitter is 11% *substance*, 5% *status*, 16% *style*, 10% *social*, and 56% *other*. Despite the common perception to the contrary, usage of *substance* dimensions outnumbers *status* dimensions on Twitter by two to one.

Other is so common because of how our 4S categorization interacts with other kinds of common trends that on Twitter. For instance, time words and numbers are contained prominently in several topics that are labeled *other*. The largest source of *other*, however, comes from the distribution of languages on Twitter. In particular, about half of user traffic comes from non-English speaking countries,[2] and the language in which a post is written is a powerful similarity signal across posts. The model effectively segregates usage of these languages into their own dimensions, which we manually labeled as *other*. Only once a language has enough posts will the model have enough data to subdivide by linguistic usage.

By aggregating Labeled LDA dimensions across recent posts from two Twitter accounts, we can visually contrast their language usage. Figure 2 shows a 4S analysis of 200 recent posts written by a popular celebrity (@oprah, right) and by the World Wide Web Consortium (@w3c, left). In the center, we see the ratios of these two account's usage of dimensions that fall into each 4S category, denoted as stacked vertical segments drawn to scale. Background statistics for the dataset are shown as a third stacked bar in the center, from which we can see that @w3c is highly skewed toward *substance*, whereas @oprah has slightly more *status* than average. The most common words for selected dimensions within each 4S category are shown to the left and right. The size of a word reflects how important it is in that dimension globally (i.e. in the training data), and shading depends upon how often the poster uses each word within that dimension.

[2] While we could not find an exact statistic for the distribution of languages by post on Twitter, English-speaking countries make up about 49% of user traffic (http://www.alexa.com/siteinfo/twitter.com).

Images like Figure 2 can be used to visually characterize and contrast users. For instance, we can see that @oprah posts about her television show (top right) and about books (adjacent in region). In particular, we see that @oprah uses the "book" dimension to talk about reading (darker) rather than writing (unshaded). Similarly, @w3c often posts about technology (top left) and the web (adjacent). Within the web topic, @w3c uses words like "internet" and "online" but not "marketing" or "seo." Socially, @w3c comes across as an open organization by using words like *join, we, our* and *us*, whereas @oprah talks to her followers (*your, you're*). A scalable, interactive version of this visualization is in development to be released on the web.

Ranking Experiments

The previous section demonstrated ways we can use Labeled LDA with a 4S analysis to characterize sets of posts according to the model's learned dimensions. Here we examine the model from a different perspective: effectiveness at modeling Twitter content as measured by performance on two information consumption tasks. One task considers ranking posts from a person's current feed; the other is aimed at recommending new users to follow. In these experiments, we do not make use of the 4S categorization of the L-LDA dimensions, instead focusing on the relative effectiveness of two representations of Twitter content: the per-post feature space defined by Labeled LDA's per-post θ_d and standard tf-idf feature vectors built from tokenized posts. We also report the performance of a combination of these models and two baseline methods, ordering randomly and ordering by time. The Labeled-LDA model used here was a 100 latent dimension model with all labeled dimensions as described above, trained on the OneWeek dataset.

Active Twitter users within Microsoft were asked to rate the quality of posts from users they follow on a three point scale. For each participating *rater*, we selected up to seven *posters* with public feeds followed by that rater. We collected the 14 most recent posts from each poster using Twitter's public API. This collection of 7×14 posts was presented to the rater in chronological order. Each rater was asked to score the selected posts on a three point scale: 3 = "must read," 2 = "maybe worth the reading time," and 1 = "not really worth reading." 43 users completed at least 60 judgments, providing us a dataset of 4,267 judgments. Most raters in our study were unhappy with most posts in their feeds. The average rating was only 1.67, with a majority of posts (2,187) scored as "not really worth reading." Individual raters displayed a range of satisfaction: the median per-rater average score was 1.64, with a minimum of 1.08 and a max of 2.26.

By-Rater Post Ranking Task

The by-rater post ranking task models a content-driven information consumption scenario: given only a few minutes, which posts should *@jane* read from her feed. To evaluate this task, we split the set of judgments by rater, ordering posts chronologically. The earliest 70% of posts were taken as a training set, and the remaining were scored as a test set, with the goal of ranking the most preferred posts first. While a more involved supervised classification algorithm could be used, here we trained a simple centroid-based ranker on the positive examples (those rated as "must read" or "maybe worth the reading time") in order to compare feature spaces. Test posts were ordered by their cosine similarity to the mean feature vector of the positive examples.[3]

Table 3: Performance on the by-rater post ranking task.

Model	Mean Average Precision	Mean Prec@1	Mean RR@1R
L-LDA + tf-idf (best)	**.622**	**.634**	**.756**
L-LDA	.604	.537	.681
tf-idf	.608	.585	.718
Temporal	.565	.537	.678
Random	.542	.537	.670

Table 4: Performance on the user recommendation task.

Model	Reciprocal Rank
L-LDA + tf-idf (best)	**.965**
L-LDA	.579
tf-idf	.839
Temporal	.103
Random	.314

Table 3 shows the results of computing several standard IR rank evaluations (Mean Average Precision, Mean Precision @ 1, and Mean Reciprocal Rank of the first relevant item) on the resulting test sets. We compared performance for models based on raw tf-idf features computed on terms in the posts, the lower dimensional feature space of Labeled-LDA, a combination of the two, a random baseline, and a baseline based on time (the Twitter default). We observe that the tf-idf and Labeled-LDA models have similar performance, but that a weighted combination of their similarity scores (18% L-LDA, 82% tf-idf) outperforms all models by a substantial margin. While a full exploration of combinations of similarity models is outside the scope of this paper, this particular mixture was picked by examining performance on a set of bootstrap samples on a fraction of our dataset; performance was fairly stable and nearly optimal across a range of values between 15% and 20% L-LDA.

User Recommendation Task

The user recommendation task models a different content-driven information need: given posts from users I follow,

[3] For the probabilistic models, we also experimented with information theoretic measures like KL-divergence, but found them inferior to cosine similarity.

recommend a new user to follow. In this task, we ignore the positive and negative per-post ratings, and simply model the centroids of posts from the rater's followed users. For each rater, we build a representation of their interests using posts from six of the posters that they follow, and hold out posts from the one remaining poster as a positive test example. As negative test examples we use 8 other posters that the rater does not follow. Models are compared by the extent to which they recommend the positive test user over the negative users. Specifically, we measure the reciprocal rank of the positive test example in the set of test posters. This measure is somewhat conservative since the rater may actually be interested in some people whom they don't currently follow, particularly because our negative test examples were drawn from within the same post ranking dataset. Because all raters work for the same company and share some interest in social networking, we expect there to be more similarity between followed users and non-followed users in this dataset than for Twitter as whole.

Table 4 shows the performance across the same models as the previous experiment. Here, the temporal baseline ranks users by their average post time, so users who posted more recently more often are ranked higher. In this task, tf-idf greatly outperforms L-LDA alone, but the combination substantially outperforms either model individually. And more pointedly, the combination classifier returns a nearly perfect score of .96 – i.e. it ranks the actually followed user first in almost all test instances.

In both tasks, the best classifier was a weighted combination of these inputs. The weighted combination works well because Labeled LDA and the tf-idf model capture different aspects of textual similarity. In particular, we expect L-LDA to outperform tf-idf when document vectors share few terms in common because L-LDA reduces the dimensionality of the word space to a much smaller label space. Conversely, we expect the tf-idf model to outperform L-LDA when there are enough terms in common such that the occasionally spurious conflations in the reduced space do more harm than good. Because both of these similarity signals are informative, the weighted combination allowed the models to complement each other and outperform either model on its own.

Conclusion

This work argues that better representations of textual content on Twitter are important for solving two categories of unmet information needs: improving methods for finding/following new users and topics, and for filtering feeds. Latent variable topic models like Labeled LDA provide a promising avenue toward solving these challenges. We have shown how these methods can support rich analyses of Twitter content at large scale and at the level of individual users with 4S analyses, mapping sets of posts into *substance, status, social,* and *style* dimensions. And we have shown how the topic models' lower dimensional feature representation can be used to characterize users by the topics they most commonly use. The approach effectively models important similarity information in posts, improving performance on two concrete tasks modeled after information needs: personalized feed re-ranking and user suggestion.

We are interested in building richer applications of Labeled-LDA and similar models of Twitter content. With larger sets of judgments and users, we can evaluate and tune more model parameters (e.g., number of topics, strategies for mixing of latent and labeled topics) and richer models for ranking and classification. Prototype interfaces are under development to support improved finding, following and search on Twitter. In future work, we plan to examine the temporal dynamics and resolution of learned topic models and to combine our new content analysis techniques with basic reputation and social network analysis. Such extensions would enable us to answer questions like: How much does the distribution of *substance, status, social,* and *style* change across parts of the social network? How does each person's usage of language evolve over time? While this work takes only a first step in richer content-based analysis of Twitter, we believe there is a bright future for such models on microblogs moving forward.

References

Asuncion, A., Smyth, P., & Welling, M. (2008). Asynchronous distributed learning of topic models. *NIPS 2008.*

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On Smoothing and Inference for Topic Models. *UAI 2009*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*.

boyd, d., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *HICSS 2010.*

Honeycutt, C., & Herring, S. (2009). Beyond microblogging: Conversations and collaboration via Twitter. *HICSS 2009.*

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we Twitter: Understanding Microblogging Usage and Communities. *WebKDD/SNA-KDD 2007.*

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). A few chirps about Twitter. *WOSP 2008.*

Naaman, M., Boase, J., & Lai, C.-H. (2010). Is it Really About Me? Message Content in Social Awareness Streams. *CSCW 2010.*

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. *EMNLP* 2009.

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised Modeling of Twitter Conversations. *NAACL* 2010.