Improving LDA Topic Models for Microblogs via Tweet Pooling

Author 1
Affiliation 1
Address 1
Country 1
email1

ABSTRACT

Twitter: the world of 140 characters poses serious challenges to the efficacy of topic models on short, messy text. While topic models such as Latent Dirichlet Allocation (LDA) have a long history of successful application to news articles and academic abstracts, they are often less coherent when applied to microblog content like Twitter. In this paper, we investigate methods to improve topics learned from Twitter content *without* modifying the basic machinery of LDA; we achieve this through various pooling schemes that aggregate tweets in a data preprocessing step for LDA. We empirically establish that tweet pooling by hashtags leads to a vast improvement in a variety of measures of topic coherence across three diverse Twitter datasets in comparison to an unmodified LDA baseline and a variety of pooling schemes.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous
; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

The "undirected informational" search task, where people seek to better understand the information available in document corpora, uses techniques such as multidocument summarisation and topic modeling. Topic models uncover the salient patterns of a collection under the mixed-membership assumption: each document can exhibit multiple patterns to different extents. When analysing text, these patterns are represented as distributions over words, called *topics*. Probabilistic topic models such as Latent Dirichlet Allocation (LDA) [1] are a class of Bayesian latent variable models that have been adapted to model a diverse range of document genres.

Table 1: Sample Topic Words

Poor Topics	Coherent Topics
barack cool apple health iphone los barackobama video uma gop	flu swine news pandemic health death flight h1n1 vaccine confirmed

We consider the application of LDA to Twitter content, which poses unique challenges different to much of standard NLP content: (1) posts are short (140 characters or less), (2) mixed with contextual clues such as URLs, tags, and Twitter names, and (3) use informal language with misspelling, acronyms and nonstandard abbreviations (e.g. O o haha wow). Hence, effectively modeling content on Twitter requires techniques that can readily adapt to this unwieldy data while requiring little supervision.

Unfortunately, it has been found that topic modeling techniques like LDA do *not* work well with the messy form of Twitter content [16]. Topics learned from LDA are formally a multinomial distribution over words, and by convention the top-10 words are used to identify the subject area or give an interpretation of a topic. The naïve application of LDA to Twitter content produces mostly incoherent topics. Table 1 demonstrates poor topic words as compared to topic words which are much more coherent and interpretable.

How can we extract better topics in microblogging environments with standard LDA? While lingistic "cleaning" of text could help somewhat, for instance [3], a complementary approach using LDA is also needed because there are so few words in a tweet. An intuitive solution to this problem is tweet pooling [13, 4]: merging related tweets together and presenting them as a single document to the LDA model. In this paper we examine various tweet-pooling schemes to improve LDA topic quality. We compare the performance of these methods across three datasets constructed to be representative of the diverse collections of content possible in the microblog environment and examine a variety of topic coherence evaluation metrics including the ability of the learned LDA topics to reconstruct known clusters and the interpretability of these topics via statistical information measures. Overall, we find that the novel method of pooling tweets by hashtags yields superior performance for all metrics on all datasets, thus providing a scheme for significantly improved LDA topic modeling on Twitter.

2. TWEET POOLING FOR TOPIC MODELS

The goal of this paper is to obtain better LDA topics from Twitter content without modifying the basic machinery of standard LDA. As noted in Section 1, microblog messages differ from conventional text: message quality varies greatly, from newswire-like utterances to babble. To address the challenges, we present various pooling schemes to aggregate tweets into "macro-documents" for use as training data to build better LDA models. The motivation behind tweet pooling is that individual tweets are very short (≤ 140 characters) and hence treating each tweet as an individual document does not present adequate term co-occurence data within documents. Aggregating tweets which are similar in some sense (semantically, temporally, etc.) enriches the content present in a single document from which the LDA can learn a better topic model. We next describe various tweet pooling schemes.

Basic scheme – Unpooled Tweets: The default treats each tweet as a single document and trains LDA on all tweets. This serves as our baseline for comparison to pooled schemes.

Author-wise Pooling: Pooling tweets according to author is a standard away of aggregating Twitter data [13, 4] and shown to be superior to unpooled Tweets. To use this method, we build a document for each author which combines all tweets they have posted.

Burst-score wise Pooling: A *trend* on Twitter [7] (sometimes referred to as a trending topic) consists of one or more terms and a time period, such that the volume of messages posted for the terms in the time period exceeds some expected level of activity. In order to identify trends in Twitter posts, unusual "bursts" of term frequency can be detected in the data. We run a simple burst detection algorithm to detect such trending topics and aggregate tweets containing those terms having high burst scores. To identify terms that appear more frequently than expected, we will assign a score to terms according to their deviation from an expected frequency. Assume that M is the set of all messages in our tweets dataset, R is a set of one or more terms (a potential trending topic) to which we wish to assign a score, and $d \in D$ represents one day in a set D of days. We then define $M(R, d)$ as the subset of Twitter messages in M such that (1) the message contains all the terms in R and (2) the message was posted during day d. With this information, we can compare the volume in a specific day to the other days. Let $Mean(R) = \frac{1}{|D|}\Sigma_{d \in D} M(R, d)$. Correspondingly, $SD(R)$ is the standard deviation of $M(R, d)$ over the days $d \in D$. The *burst-score* is then defined as:

$$burst\text{-}score(R, d) = \frac{|M(R, d) - Mean(R)|}{SD(R)}$$

Let us denote an individual term having burst-score greater than some threshold τ on some day $d \in D$ as a *burst-term*. Then our first novel aggregation method of Burst Score-wise Pooling aggregates tweets for each burst-term into a single document for training LDA, where we found $\tau = 5$ to provide best results.

Temporal Pooling: When a major event occurs, a large number of users often start tweeting about the event within a short period of time. To capture such temporal coherence of tweets, the fourth scheme and our second novel pooling proposal is known as Temporal Pooling, where we pool all tweets posted within the same hour.

Hashtag-based Pooling: A Twitter *hashtag* is a string of characters preceded by the hash (#) character. In many cases hashtags can be viewed as topical markers, an indication to the context of the tweet or as the core idea expressed in the tweet, therefore hashtags are adopted by other users that contribute similar content or express a related idea. One example of the use of hashtags is "ask GAGA anything using the tag #GoogleGoesGaga for her interview! RT so every monster learns about it!! " referring to an exclusive interview for Google by Lady Gaga (singer). For the hashtag-based pooling scheme, for each create pooled documents for each hashtag. If any tweet has more than one hashtag, this tweet gets added to the tweet-pool of each of those hashtags.

Other Pooling: While a few other combinations of pooling schemes (eg.author-time, hashtag-time, *etc*) are possible, the initial results obtained were not as good as those presented for the currently outlined pooling schemes.

Table 2: Datasets

Dataset	Term/%
Generic	music/17.9 business/15.8 movie/14.5 design/10.8 food/9.6 fun/9.1 health/6.9 family/6.4 sport/4.9 space/3.2
Specific	Obama/23.2 Sarkozy/0.4 baseball/3.5 cricket/1.8 Mcdonalds/1.5 Burgerking/0.5 Apple/16.3 Microsoft/6.8 United-states/40.7 France/4.9
Events	Flight-447/0.9 Jackson/13.9 Lakers/13.8 attack/13.8 scandal/4.1 swine-flu/13.8 recession/12.3 conference/14.1 T20/4.4 Iran-election/8.6

3. TWITTER DATASET CONSTRUCTION

We construct three datasets representative of the diverse collections of content found on Twitter. We chose one or two term queries (often with similar pairs of queries to encourage a non-strongly diagonal confusion matrix) to search a tweet collection and each resulting set of tweets was labeled by the query that retrieved it. Since the number of queries (equivalently the number of clusters) is known beforehand, we could use this knowledge to evaluate how well the topics output by LDA match with known clusters. A brief description of the three datasets is as follows:

Generic Dataset: 359478 tweets from 11 Jan'09 to 30 Jan'09. A general dataset with tweets containing generic terms.

Specific Dataset: 214580 tweets from 11 Jan'09 to 30 Jan'09. A dataset composed of tweets which have specific terms that refer to specific named entities.

Event Dataset: 207128 tweets from 1 Jun'09 to 30 Jun'09. A dataset composed of tweets pertaining to specific events. The query terms represent these events and the time period was chosen specifically due to the number of co-occurring events being discussed at this time.

Table 2 provides the exact query terms and the percentage of tweets in the datasets retrieved by each query. Typically, less than one percent of tweets were retrieved by more than one query with the highest case of 4.6% overlap occurring in the *generic dataset* for the two queries "family" and "fun". We have removed tweets retrieved by more than one query in a dataset in order to preserve uniqueness of tweet labels for later analysis with clustering metrics.

4. EVALUATION METRICS USED

Because there is no single method for evaluating topic models, we evaluate a range of metrics including those used in clustering (purity and NMI), semantic topic coherence and interpretability (PMI), and a pure probabilistic approach (held-out probability) [12].

In order to cluster with LDA, we let a topic represent each cluster and assign each tweet to its corresponding mixture topic of highest probability (an inferred quantity via LDA). Then by analysing clustering-based metrics, we wish to understand how well the different tweet pooling schemes are able to cluster according to the original queries used to produce the datasets.

Formally, let T_i be the set of tweets in LDA topic cluster i and Q_j be the set of tweets with query label j. Then let $T = \{T_1, \ldots, T_{|T|}\}$ be the set of all $|T|$ clusters and $Q = \{Q_1, \ldots, Q_{|Q|}\}$ be the set of all $|Q|$ query labels. Now we define our clustering-based metrics as follows.

Purity: To compute purity [5], each LDA topic cluster is assigned the *query label most frequent in the cluster*. Purity then simply measures the average "purity" of each cluster, i.e., the fraction of tweets in a cluster having the assigned cluster query label. Formally:

$$Purity(T, Q) = \frac{1}{|T|}\Sigma_{i \in \{1 \ldots |T|\}} max_{j \in \{1 \ldots |Q|\}} |T_i \cap Q_j|$$

Obviously, high purity scores reflect better original cluster reconstruction.

Normalized Mutual Information (NMI): As a more information-theoretic measure of cluster quality, we also evaluate normalized mutual information (NMI) defined as follows

$$NMI(T,Q) = \frac{2I(T;Q)}{H(T)+H(Q)},$$

where $I(\cdot,\cdot)$ is mutual information and $H(\cdot)$ is entropy as defined in [5]. NMI is always a number between 0 and 1 and will be 1 if the clustering results exactly match the category labels while 0 if the two sets are independent.

Learnt topics should be coherent and interpretable. Topic coherence – meaning semantic coherence – is a human-judged quality that depends on the semantics of the words, and cannot be measured by model-based statistical measures that treat the words as exchangeable tokens. It is possible to automatically measure topic coherence with near-human accuracy [10] using a score based on pointwise mutual information (PMI). We use this to measure coherence of the topics from different tweet-pooling schemes.

Pointwise Mutual Information (PMI): PMI is a measure of the statistical independence of observing two words in close proximity where the PMI of a given pair of words (u,v) is $PMI(u,v) = log\frac{p(u,v)}{p(u)p(v)}$. Both probabilities are determined from the empirical statistics of the full collection, and we treat two words as co-occurring if both words occur in the same tweet.

For a topic T_k ($k \in \{1 \dots |T_k|\}$), we measure topic coherence as the average of PMI for the pairs of its ten highest probability words $\{w_1, ..., w_{10}\}$:

$$PMI\text{-}Score(T_k) = \frac{1}{100}\Sigma_{i=1}^{10}\Sigma_{j=1}^{10}PMI(w_i,w_j).$$

The average of the PMI score over all the topics is used as the final measure of the PMI score.

Held-out Probability: A final way of evaluating topic models is to compare predictive performance by estimating the probability of a subset of held-out documents. We used the “Left to Right” evaluation algorithm as described in [12] to calculate these values, which is an unbiased method. Another approach is the so-called document completion method [12], however with so few words we felt holding out a subset of a (small) document was ill-advised.

5. RESULTS FOR POOLING SCHEMES

In this section we discuss the results of the experimental evaluation of the tweet pooling schemes introduced in Section 2. The datasets used were described in Section 3 while the evaluation metrics used were described in Section 4.

5.1 Document Characteristics

We first have a look at the document characteristics of the documents in the different pooling schemes for the three datasets as they may affect LDA. Table 3 presents these statistics.

The statistics presented above highlight the differences in the characteristics of the documents on which LDA models have been trained. The number of documents decreases as we move from Unpooled scheme to Authorwise and Hashtagwise pooling scheme, while the corresponding size of the documents in each case increases. On an average the document size increases by a factor of seven in hashtag-based pooling when compared against unpooled or authorwise pooling schemes. On the other extreme lies the temporal pooling with very less number of documents and hence each document of a much larger size. Some level of pooling is good, but it should also retain semantic coherence.

5.2 Comparison of Pooling Schemes

For the three datasets (viz. Generic, Specific and Events) and pooling schemes, we next evaluate the Purity scores, NMI scores, PMI scores and the Held-out probabilities in Table 4 on the topic model obtained by training LDA using each scheme.

Based on these results we conclude that hashtag-based pooling scheme *clearly* performs better than unpooled scheme as well as other pooling schemes.

6. RELATED WORK

Our work is quite different from many pioneering studies on topic modeling of Tweets because we focus on how we could improve topic coherence with existing algorithms. Prior work on topic modeling for tweets includes the work of [11] which presents a scalable implementation of a partially supervised learning model (Labeled LDA). [16] empirically compare the content of Twitter with a traditional news medium, New York Times, using unsupervised topic modeling. [4] use the topic modeling approach for predicting popular Twitter messages and classifying Twitter users and corresponding messages into topical categories. [**?**] propose a novel method for normalising ill-formed out-of-vocabulary words in short microblog messages. The TwitterRank system [13] and [4] uses author-based pooling to apply LDA to tweets. [16] compared topic characteristics between twitter and traditional news media; they propose to use one topic per tweet (similar to PLSA), and argue that this is better than no pooling, or the author-topic model. [9] proposed two methods to regularize the resulting topics towards better coherence. sparseness in LDA, the weights are derived from LSA vector length. [8] used LDA for tweet retrieval. In addition, they used retweet as an indicator of "interestingness" to improve retrieval quality. This related research suggests a number of orthogonal methods that could be used to complement our tweet pooling scheme.

7. CONCLUSION

This paper presents a way of aggregating tweets in order to improve performance of topic models in terms of quality of topics obtained measured by the ability to reconstruct clusters and topic coherence. The results presented in Table 4 suggest that hashtag-based pooling outperforms all other pooling strategies including the default way of training topic models on unpooled Twitter data.

8. REFERENCES

[1] D. Blei, A. Ng, and M. Jordon. Latent Dirichlet allocation. volume 3, pages 993–1022, 2003.

[2] B. Han, P. Cook, and T. Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proc. of EMNLP-CoNLL 2012*, pages 421–432, Korea, 2012.

[3] B. Han, P. Cook, and T. Baldwin. Lexical normalisation of short text messages. *ACM Transactions on Intelligent Systems and Technology*, to appear, 2012.

[4] L. Hong and B. Davison. Empirical study of topic modeling in Twitter. In proceedings 1st ACM Workshop on Social Media Analytics, 2010.

[5] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[6] M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on Twitter. In proceedings Journal of the American Society for Information Science and Technology (JASIST), 62(5). doi: 10.1002/asi.21489, 2011.

Table 3: Document Characteristics for different schemes

Pooling Scheme	#of docs			Avg # of words/doc			Max # of words/doc		
	Generic	Specific	Events	Generic	Specific	Events	Generic	Specific	Events
Authorwise	208300	118133	67387	17.6	20.4	15.4	4893	3586	2775
Unpooled	359478	214580	207128	10.2	10.9	9.7	35	49	32
Burst Score	7658	7436	5434	76.5	154.2	71.6	61918	420249	57794
Hourly	465	464	463	8493.4	5387.5	2422	20144	18869	38893
Hashtag	8535	7029	4099	70.4	187.2	78.4	61918	420249	57794

Table 4: Results of different pooling schemes

Scheme	Purity			NMI Score			PMI score			Log Held-out Probability		
	Generic	Specific	Events	Generic	Specific	Events	Generic	Specific	Events	Generic	Specific	Events
Unpooled	0.49 ± 0.08	0.64 ± 0.07	0.69 ± 0.09	$\mathbf{0.28 \pm 0.04}$	0.22 ± 0.05	0.39 ± 0.07	-1.27 ± 0.11	0.47 ± 0.12	0.47 ± 0.13	-82.2 ± 6.3	-89.3 ± 7.2	-86.3 ± 7.4
Author	$\mathbf{0.54 \pm 0.04}$	0.62 ± 0.05	0.60 ± 0.06	0.24 ± 0.04	0.17 ± 0.04	0.41 ± 0.06	0.21 ± 0.09	0.79 ± 0.15	0.51 ± 0.13	-63.0 ± 4.3	-68.6 ± 4.7	-66.4 ± 5.2
Hourly	0.45 ± 0.05	0.61 ± 0.06	0.61 ± 0.07	0.07 ± 0.04	0.09 ± 0.04	0.32 ± 0.05	-1.31 ± 0.12	0.87 ± 0.16	0.22 ± 0.14	-64.8 ± 6.2	-69.4 ± 5.8	-67.9 ± 7.1
Burstwise	0.42 ± 0.07	0.60 ± 0.04	0.64 ± 0.06	0.18 ± 0.05	0.16 ± 0.04	0.33 ± 0.04	0.48 ± 0.16	0.74 ± 0.14	0.58 ± 0.16	-56.7 ± 5.5	-59.0 ± 4.5	-57.8 ± 6.1
Hashtag	$\mathbf{0.54 \pm 0.04}$	$\mathbf{0.68 \pm 0.03}$	$\mathbf{0.71 \pm 0.04}$	$\mathbf{0.28 \pm 0.04}$	$\mathbf{0.23 \pm 0.03}$	$\mathbf{0.42 \pm 0.05}$	$\mathbf{0.78 \pm 0.15}$	$\mathbf{1.43 \pm 0.14}$	$\mathbf{1.07 \pm 0.17}$	**-55.9 ± 4.3**	**-58.9 ± 4.1**	**-55.4 ± 4.3**

[7] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proceedings of CIKM '11*, pages 183–188, New York, NY, USA, 2011. ACM.

[8] D. Newman, E. Bonilla, and W. Buntine. Improving topic coherence with regularized topic models. In proceedings of NIPS, 2011.

[9] D. Newman, J. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In proceedings NAACL, 2010.

[10] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In proceedings AAAI Conference on Weblogs and Social Media, 2010.

[11] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In proceedings ICML, 2009.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings WSDM '10*, pages 261–270, New York, NY, USA, 2010. ACM.

[13] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of ECIR'11*, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.