Dynamic Multi-Faceted Topic Discovery in Twitter

Jan Vosecky, Di Jiang, Kenneth Wai-Ting Leung, Wilfred Ng
Hong Kong University of Science and Technology
Kowloon, Hong Kong, China
{jvosecky,dijiang,kwtleung,wilfred}@cse.ust.hk

ABSTRACT

Microblogging platforms, such as Twitter, already play an important role in cultural, social and political events around the world. Discovering high-level topics from social streams is therefore important for many downstream applications. However, traditional text mining methods that rely on the bag-of-words model are insufficient to uncover the rich semantics and temporal aspects of topics in Twitter. In particular, topics in Twitter are inherently dynamic and often focus on specific entities, such as people or organizations. In this paper, we therefore propose a method for mining multi-faceted topics from Twitter streams. The Multi-Faceted Topic Model (MfTM) is proposed to jointly model latent semantics among terms and entities and captures the temporal characteristics of each topic. We develop an efficient online inference method for MfTM, which enables our model to be applied to large-scale and streaming data. Our experimental evaluation shows the effectiveness and efficiency of our model compared with state-of-the-art baselines. We further demonstrate the effectiveness of our framework in the context of tweet clustering.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

Keywords

Twitter; Topic Model; Unsupervised Learning; Clustering

1. INTRODUCTION

In recent years, social media and in particular microblogs have seen a steep rise in popularity, with users from a wide range of backgrounds contributing content in the form of short text-based messages. Twitter, a popular microblogging platform, is at the epicenter of the social media explosion, with millions of users being able to create and publish short posts, referred to as *tweets*, in real time. Discovering high-level topics from social streams is therefore important for many downstream applications, such as classification, clustering and user modeling [10, 13, 15].

However, there is still a lack of accurate and efficient models for automatic topic discovery in microblogs. In contrast to traditional domains such as news documents or scientific literature, topic discovery in microblogs faces many new challenges. We summarize the main challenges as follows.

Entity-centric. Microblog posts often discuss specific entities, such as famous people, organizations, or geographic locations [2]. Traditional models for textual data based on the vector-space model or topic modeling take a simplistic bag-of-words view of a "topic" [1, 5, 13]. These methods fail to distinguish the rich semantics of microblog topics and exploit the various entity types.

Highly dynamic: Microblog topics are constantly evolving, implying the need to model their temporal characteristics. However, the temporal dimension has not been sufficiently explored in current topic models for Twitter data. Moreover, the real-time and streaming nature of social content calls for scalable and updatable models.

Figure 1 illustrates the mentioned characteristics of a topic in Twitter.

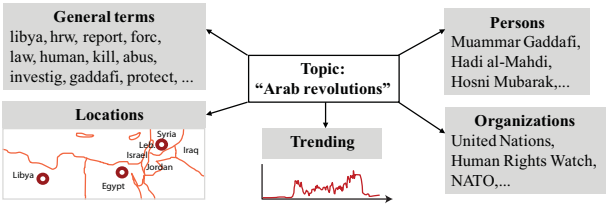


Figure 1: Multiple facets of a topic discussed in Twitter

To tackle these issues in a unified way, we propose a novel topic discovery method. At the core, we propose the Multi-Faceted Topic Model (MfTM), which extracts rich latent topics from microblog content and the associated temporal patterns. In essence, each latent topic has multiple orthogonal 'facets'. For example, the latent topic 'Arab revolutions' may consist of five facets: general terms (e.g. 'libya', 'war', 'protest'), person names (e.g. 'Muammar Gaddafi'), organizations (e.g. 'United Nations'), location names (e.g. 'Libya', 'Egypt') and a temporal distribution, indicating the trending behavior of the topic. As we show in this paper, the MfTM is more suitable to social streams than standard topic models, such as LDA.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505593.

Parameter inference is a known bottleneck of topic models, in particular in face of the scale of microblog data. We therefore build upon the recent advances in variational inference methods and develop an "online" inference algorithm for MfTM. In contrast to Gibbs sampling or batch variational inference, our algorithm processes data sequentially. As we show in our evaluation, our inference method easily scales to large datasets and has the advantage of continuous updatability.

The performance of our topic discovery method is thoroughly evaluated and compared against multiple baseline methods. On the task of tweet clustering, we demonstrate the benefits of our model for downstream applications.

The rest of the paper is structured as follows: Section 2 reviews related work. In Section 3, we present our topic discovery method. Section 4 presents our experimental evaluation. We conclude our findings in Section 5.

2. RELATED WORK

Our topic modeling approach is related to previous works on probabilistic topic models. Blei et al. [1] proposed Latent Dirichlet Allocation (LDA) to analyze electronic archives. Topic models were since applied in various domains, including search query logs [6, 7] and app marketplaces [8]. In the microblogging environment, Hong et al. [5] study approaches to apply LDA on microblog data. Topic models are used in [18] to compare topics in Twitter and in news articles. In [16], a topic modeling approach is used to discover geographic user interests.

To enrich the bag-of-words representation of a topic, some models were proposed to consider additional semantics. The topic-aspect model by Paul and Girju [12] is proposed to model "multi-faceted" topics. In their definition, a "multi-faceted" topic is a topic that is expressed differently across different aspects, such as scientific disciplines. Our focus and definition of multi-faceted topics is therefore fundamentally different. The entity-topic model by Newman et al. [11] considers two facets of a topic: general terms and entities. In contrast, our model considers general terms and each entity type in a separate facet, as well as a temporal facet. In [17], the timestamps of a topic are assumed to follow a beta distribution. In contrast, we model timestamps as a multinomial distribution, thus enabling our model to capture arbitrary temporal patterns.

Our work also builds upon recent advances in topic model inference, in particular *stochastic variational inference* (SVI) [4]. SVI enables topic models to be trained on massive and streaming data, since it operates in a sequential rather than batch fashion (such as Gibbs sampling). We adopt this technique to develop an online learning method for MfTM.

Our approach shows a new direction in the use of topic models in social media. This is to the best of our knowledge the first work that proposes to organize the interplay between general terms, entities and time in a principled manner. As shown in our experiments, our method can consistently outperform standard LDA.

3. TOPIC DISCOVERY IN TWITTER

In this section, we present our method for topic discovery from Twitter streams. We first discuss pre-processing steps. Then we present the novel Multi-Faceted Topic Model and its online inference algorithm.

3.1 Data Pre-processing

3.1.1 Data Normalization

Due to the informal nature of microblog posts, a number of cleansing steps are performed. First, posts are converted to lower-case, punctuation and numbers are removed and characters repeated consecutively more than twice are stripped, in order to correct basic misspellings (e.g. the string "goooood" will be converted to "good"). Second, URL links are stored separately for further use and removed from the post. Third, stopwords are removed and all terms are stemmed using a standard Porter stemmer.

3.1.2 Entity Extraction

URL links contained in posts provide an opportunity to obtain additional semantics from the referred web documents. Specifically, we utilize the referred web documents to extract named entities. We choose this approach over other methods, such as matching Wikipedia entries [10], since our approach does not require a matching algorithm, thus reducing computational cost. We first follow the URLs mentioned in tweets and crawl the web documents. Second, we perform named entity recognition (NER) using the Stanford NER library[1]. In general, our framework is able to accommodate an arbitrary number of named entity types. In this paper, we focus our attention on 'person', 'organization' and 'location' named entities. Apart from web documents, we note that named entities can also be extracted directly from microblog posts. However, the short and informal nature of microblog content results in a poor accuracy of conventional NER tools [14]. In principle, tweet-based named entity extraction can be seamlessly integrated into our topic discovery method with the availability of appropriate NER tools.

3.2 Multi-Faceted Topic Model

Traditional topic models, such as Latent Dirichlet Allocation (LDA) [1], can be used to learn a set of latent topics. Each topic in LDA is a multinomial distribution over words. In contrast, we aim to model latent topics with finer granularity, such as preference towards specific entities (e.g., specific locations) and the topic's temporal characteristics. We therefore propose the *Multi-Faceted Topic Model* (MfTM) to discover rich latent topics from Twitter data.

In MfTM, we assume the existence of X types of elements in the microblog corpus. In this paper, we focus on five particular element types. First, we distinguish three named entity types, comprising *person* (e_p), *organization* (e_o) and *location* (e_l) entities. Similarly to LDA, MfTM also models general terms, which are treated as *term* elements (e_t). Additionally, we capture the trending behavior of topics over time. Tweet timestamps are discretized into fixed-length intervals and treated as *time* elements (e_τ). In general, the length of the time interval depends on the desired temporal granularity. In our work, timestamps are discretized into day-intervals, since a day is a commonly used unit for grouping news-related content.

In MfTM, elements of each type follow a multinomial distribution given a latent topic. In other words, each element type forms an orthogonal *facet* of a latent topic. Figure 2 illustrates the structure of the model.

The generative process of MfTM proceeds as follows:

[1]http://nlp.stanford.edu/ner/

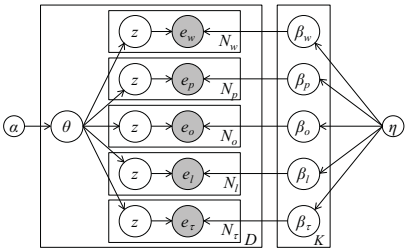


Figure 2: Graphical model of MfTM

1. For each topic $k \in \{1, \ldots, K\}$:
   - For each facet $x \in \{1, \ldots, X\}$:
     - Draw facet x of topic k: $\beta_x^k \sim \mathrm{Dir}_{V_x}(\eta)$.
2. For each document $d \in \{1, \ldots, D\}$:
   - Draw document's topic mixture $\theta_d \sim \mathrm{Dir}_K(\alpha)$.
   - For each facet $x \in \{1, \ldots, X\}$ and element position $n \in \{1, \ldots, N_{d,x}\}$:
     - Draw topic assignment $z_{d,x,n} \sim \mathrm{Mult}_K(\theta_d)$.
     - Draw element $e_{d,x,n} \sim \mathrm{Mult}_{V_x}(\beta_x^z)$.

Given the hyperparameters α and η, the joint distribution of topics β, document-topic mixture θ, topic assignments $\mathbf{z}$ and elements $\mathbf{e}$ is given by:

$$p(\theta, \beta, \mathbf{z}, \mathbf{e}|\alpha, \eta) = p(\theta|\alpha)p(\beta|\eta)\prod_{x=1}^{X} p(\mathbf{z_x}|\theta)p(\mathbf{e_x}|\mathbf{z_x}, \beta_x), \tag{1}$$

where $\mathbf{e_x}$ are all elements of type x in the corpus and $\mathbf{z_x}$ are the topic assignments of all elements of type x.

Since exact inference for this model is intractable, an approximate posterior inference method is needed to estimate the latent parameters. Although Gibbs sampling is a widely adopted inference method for topic models, an online learning method for LDA, namely *stochastic variational inference* (SVI), has been developed recently [4]. In stochastic optimization, we find the maximum of the variational objective by following noisy estimates of its natural gradient. SVI enables parameter inference on massive and streaming data, since it operates in a sequential, rather than batch fashion. This inference method fits well in the scenario of analyzing microblog posts, which essentially arrive in a streaming fashion. We use SVI as a basis to develop an online learning method for MfTM.

3.2.1 Online Inference for MfTM

We now proceed to present our online inference algorithm for MfTM. Due to space constraints, we only present the major components of the algorithm. Interested readers may refer to [4] for full details of stochastic variational inference. We begin by listing the complete conditionals of the model.

Local hidden variables. The complete conditional of the topic assignment $z_{d,x,n}$ of entity $e_{d,x,n}$ is a multinomial,

$$p(z_{d,x,n} = k|\theta_d, \beta_x^k, e_{d,x,n}) \propto \exp\{\log\theta_d^k + \log\beta_{x,e_{d,x,n}}^k\}. \tag{2}$$

The complete conditional of document d's topic distribution is a posterior Dirichlet,

$$p(\theta_d|\beta, \mathbf{z_d}) = \mathrm{Dir}(\alpha + \sum_{x=1}^{X}\sum_{n=1}^{N_{d,x}} z_{d,x,n}), \tag{3}$$

where $\mathbf{z_d}$ are the topic assignments of all elements in d.

Algorithm 1 Stochastic Variational Inference for MfTM

```
Initialize λ^(0) randomly.
Initialize γ^(0) = α.
repeat
    Sample a document d from the data set.
    Initialize intermediate local topic proportion γ̂_d = θ_d.
    repeat
        for x ∈ {1,...,X}, n ∈ {1,...,N_{d,x}} do
            for k ∈ {1,...,K} do
                Set φ^k_{d,x,n} ∝ exp{E[log θ^k_d] + E[log β^k_{x,e_{d,x,n}}]}.
            end for
        end for
        γ_d = α + Σ_{x=1}^{X} Σ_{n=1}^{N_{d,x}} φ_{d,x,n}.
    until γ_d converges
    for x ∈ {1,...,X} do
        Set intermediate topics:
        λ̂_{k,x} = η + D Σ_{n=1}^{N_{d,x}} φ^k_{d,x,n} e_{d,x,n}.
        Set global topics λ^(i+1)_{k,x} = (1 − ρ^(i)) λ^(i)_{k,x} + ρ^(i) λ̂_{k,x}.
    end for
until forever
```

Global hidden variables. The complete conditional for facet x of topic k is also a posterior Dirichlet,

$$p(\beta_x^k|\mathbf{z_x}, \mathbf{e_x}) = \mathrm{Dir}(\eta + \sum_{d=1}^{D}\sum_{n=1}^{N_{d,x}} z_{d,x,n}^k e_{d,x,n}). \tag{4}$$

The parameters of the variational distribution are chosen as follows:

- Global per-topic Dirichlets $\lambda_{1:K,1:X}$
- Local per-document Dirichlets $\gamma_{1:D}$
- Local per-word multinomials $\phi_{1:D,1:X,1:N_{d,x}}$

Each update of the local variables $\phi_{d,x,n}^k$ is defined as

$$\phi_{d,x,n}^k \propto \exp\{\mathbb{E}[\log\theta_d^k] + \mathbb{E}[\log\beta_{x,e_{d,x,n}}^k]\}. \tag{5}$$

After each update of ϕ_d, γ_d is updated as

$$\gamma_d = \alpha + \sum_{x=1}^{X}\sum_{n=1}^{N_{d,x}} \phi_{d,x,n}. \tag{6}$$

After fitting the local variables, we set the intermediate topics as

$$\hat{\lambda}_{k,x} = \eta + D\sum_{n=1}^{N_x} \phi_{d,x,n}^k e_{d,x,n}. \tag{7}$$

Finally, after processing the i^{th} document, we set the global topics as

$$\lambda_{k,x}^{(i+1)} = (1-\rho^{(i)})\lambda_{k,x}^{(i)} + \rho^{(i)}\hat{\lambda}_{k,x}, \tag{8}$$

where $\rho^{(i)} = (i + \overline{\tau})^{-\kappa}$. The parameter $\kappa \in (0.5, 1]$ is the *forgetting rate*, which controls the weight of fresh content. The *delay* $\overline{\tau} \geq 0$ is used to demote early iterations. The full inference algorithm is presented in Algorithm 1.

After applying MfTM to a training corpus, we may obtain the topic vector θ_d of a new document d as follows

$$\theta_d^k = \alpha + \prod_{x=1}^{X}\prod_{n=1}^{N_{d,x}} \beta_{x,e_{d,x,n}}^k e_{d,x,n}. \tag{9}$$

Table 1: Twitter Dataset Statistics

No. of tweets	2,126,899
No. of users	2,574
% of tweets w/ named entities (NE)	38.2%
... % of tweets with "person" NE	39.4%
... % of tweets with "organization" NE	49.8%
... % of tweets with "location" NE	29.4%

4. EVALUATION

In this section, we first describe our evaluation dataset. Second, we evaluate the proposed topic model using internal metrics, such as perplexity and topic distinctiveness. We also evaluate scalability of our inference algorithm. Third, we study the effectiveness of our framework on the task of tweet clustering.

4.1 Dataset Collection and Modeling Phase

4.1.1 Dataset Collection

To construct our evaluation dataset, we crawled publicly accessible data from Twitter using Twitter's REST API[2]. Our dataset consists of two parts: tweets by *popular users* and tweets by *general users*.

Popular users. We selected an initial set of 50 seed users from Listorius[3], a web-based service that categorizes popular Twitter users into various topical categories. The users are randomly selected from 5 different categories (technology, business, politics, celebrities and activism). Starting with these seed users, we crawled Twitter users' posts in a breadth-first search manner by traversing the followee graph. For each user, we stored up to 1,000 recent posts and selected top 20 followees of the user to add to the crawl queue. The followee selection criteria is based on the number of times the user has re-tweeted or mentioned the followee. This dataset part contains 328,428 tweets by 1,874 users in total.

General users. Twitter's Streaming API[4] provides a sample of the full public Twitter stream. We monitored the stream for one day in April 2013 and selected 700 users who posted English-language tweets and had at least 3,000 posts in total. For each of these users, we crawled up to 3,000 tweets. This dataset part contains 1,798,471 tweets in total and spans a time period from January 2009 to April 2013.

Table 1 shows the statistics of our dataset. A high-level analysis has shown that nearly 40% of tweets in our dataset mention an entity, showing that our multi-faceted model is applicable to a large proportion of tweets.

4.1.2 Modeling Phase

After collecting Twitter posts from each user, we preprocess the data as described in Section 3.1. It has been shown in [5] that grouping all posts of a user as a single document produces more accurate topic models compared with treating each post as a separate document. In our work, all user's posts published during the same day are grouped as a document. The resulting *user-day documents* thus have timestamps discretized into day-intervals.

We set the hyperparameters for MfTM in accordance with common practice in topic modeling, $\alpha = \eta = 1/K$. To select suitable values for the parameters κ and $\overline{\tau}$ in stochastic variational inference, we performed a series of experiments with $K = 50$. We vary each parameter while keeping the others fixed and observe the per-word perplexity of the model (cf. Equation 10). Finally, we set $\kappa = 0.7$ and and $\overline{\tau} = 4$.

In addition to training MfTM by means of stochastic variational inference, we also implement a Gibbs sampler for MfTM for comparison. Due to space constraints, we omit the details of the Gibbs sampling procedure. We apply the Gibbs sampler on a reduced dataset of 320,000 posts due to a longer training time required by the sampling procedure.

Figure 3 shows an example of the latent topics produced by MfTM from our dataset. To draw each topic's time distribution, we plot the multinomial values of the temporal facet in chronological order.

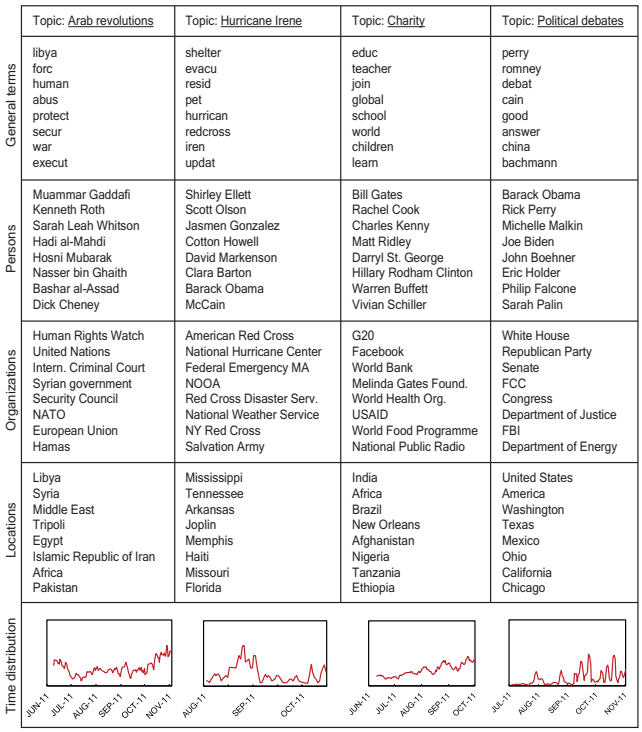


Figure 3: Example latent topics produced by MfTM. The topics titles are the authors' interpretation.

4.2 Topic Model Evaluation

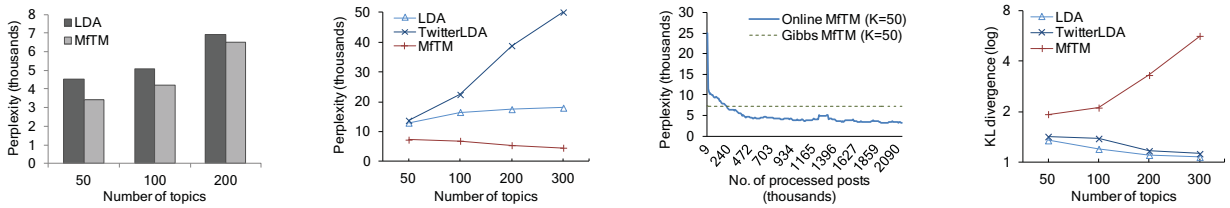
4.2.1 Perplexity Evaluation

Perplexity is a standard metric to evaluate a topic model's capability of predicting unseen data. After training the model on the training dataset, we compute the perplexity of heldout data to evaluate the models. Formally,

$$Perplexity(D_{test}|\mathcal{M}) = exp(-\frac{\sum_{d \in D_{test}} \log p(\overrightarrow{w}_d|\mathcal{M})}{\sum_{d \in D_{test}} N_d}), \quad (10)$$

where $\mathcal{M}$ is the model learned from the training dataset, $\overrightarrow{w}_d$ is the word vector for document d and N_d is the number of words in d. A lower perplexity score indicates better generalization performance of the model. As baselines for comparison, we choose LDA [1] and Twitter-LDA [18].

[2]https://dev.twitter.com/
[3]http://www.listorious.com
[4]https://dev.twitter.com/docs/streaming-apis



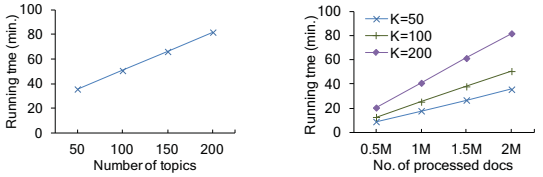
(a) Perplexity evaluation using Online inference

(b) Perplexity evaluation using Gibbs sampling

(c) Perplexity comparison of SVI and GS

(d) Topic distinctiveness

Figure 4: Topic model evaluation



(a) Topic amount

(b) Document amount

Figure 5: Scalability evaluation of online inference

Figure 4(a) presents the perplexity comparison of MfTM and LDA built using online inference, across different values of K. In Figure 4(b), we show the perplexity of MfTM, LDA and Twitter-LDA learned using Gibbs sampling (GS). From both figures, we can see that MfTM outperforms the baseline models. When GS is used to build the models (cf. Figure 4(b)), we can observe that the perplexity of LDA and Twitter-LDA increases with the number of topics, while that of MfTM decreases. These results indicate that MfTM potentially supports finer topics than the baseline models.

To illustrate the differences when using GS and our inference algorithm to train MfTM, we show the change in perplexity during the online learning of MfTM in Figure 4(c). The dotted line indicates the final perplexity after 1,000 iterations of GS on the dataset. We observe that online inference is able to reach the perplexity of the GS-learned model already after processing 200,000 posts.

4.2.2 Topic Distinctiveness

To evaluate the distinctiveness of the discovered topics, we calculate the average Kullback-Leibler (KL) divergence between each pair of topics. KL-divergence is a standard metric to evaluate the distance between two distributions. The higher the average KL-divergence, the more distinct the discovered topics are.

From the results presented in Figure 4(d), we see that the topics discovered by MfTM enjoy a much higher KL-divergence, indicating that the topics are more distinct than those discovered by LDA and Twitter-LDA. Furthermore, as the number of topics increases from 50 to 300, the average KL-divergence of topics discovered by LDA is decreasing while that of MfTM is increasing. This result again verifies our assumption that MfTM potentially supports more and finer topics than LDA.

4.2.3 Scalability

To illustrate the runtime requirements of the online inference algorithm for MfTM, we conduct a scalability evaluation. We run the experiments using a standard PC with a dual-core CPU, 4GB RAM and a 600GB hard-drive. First, we measure the time to train MfTM using different values of K and a fixed dataset size of 2 million tweets. The results in Figure 5(a) indicate a near-linear increase of training time as K increases. Second, we measure time to process a specified number of documents. Figure 5(b) illustrates that the inference algorithm is suitable for processing streaming data, since it essentially requires constant time to process each document. The timing results clearly show that SVI inference enjoys good scalability in face of voluminous data.

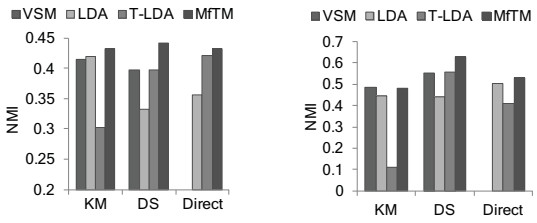
4.3 Clustering Evaluation

In this section, we evaluate the effectiveness of our model in the context of tweet clustering. Clustering tweets is a challenging task due to their short length [9, 15]. The performance of traditional text mining techniques is negatively affected in this situation, since the bag-of-words representation results in sparse instances. In contrast, our framework utilizes named entities and timestamps as additional semantic dimensions. We first describe two datasets used to conduct our clustering experiments.

Manually Labeled Dataset (ML). We invite three human reviewers and ask them to select 10 queries of their choice. For each query, we crawl the top 50 tweets returned by Twitter Search. Each reviewer is then asked to read the returned tweets and assign topic labels. Each topic label is a short free-form phrase that describes the main story of the tweet. We note that we choose to use free-form labels over a predefined taxonomy, mainly because of the diversity and evolving nature of topics in Twitter. The reviewers are asked to use a consistent set of topic labels when reviewing the list of tweets for a query. In total, we obtained 1,524 labeled tweets for 32 queries, with an average of 47.6 tweets per query. The tweets' topic labels serve as the ground truth when evaluating clustering quality. Based on the topic labels, there are 9.4 "ideal" clusters for each query on average.

Hashtag Labeled Dataset (HL). To obtain a larger dataset for comparison of clustering performance, we utilize hashtags in tweets as topic labels. We make use of the fact that Twitter users include hashtags in their tweets to indicate the tweet's topic. We first extract 100 most popular hashtags from our dataset. We then divide them into 10 batches, each batch containing 10 hashtags. For each hashtag, we select tweets containing the respective hashtag from our Twitter dataset. Before performing clustering, all hashtags are removed from the tweets. Our clustering goal in then to place tweets containing the same hashtag into the same cluster. In total, the hashtag-labeled dataset contains 7,901 tweets, each batch containing 790 tweets on average.

We use Normalized Mutual Information (NMI) as the metric for evaluating clustering quality of our labeled datasets.



(a) Dataset ML (b) Dataset HL

Figure 6: Tweet clustering evaluation on the manually-labeled (ML) and hashtag-labeled (HL) datasets

We perform clustering for each query in the ML dataset and each batch in the HL dataset and report the average NMI.

As baseline representations, we choose LDA, Twitter-LDA [18] and the vector-space model (VSM) with TF-IDF term weighting. LDA, Twitter-LDA and MfTM are built in several versions with a different number of topics (50, 100, 200). To perform tweet clustering, we use the following three algorithms:

- *K-means.* Traditional algorithm for text clustering. For the ML dataset, we set K equal to the number of unique topic labels for the respective query. For the HL dataset, K is set to the number of unique hashtags in a batch. As distance metrics, we use cosine distance for VSM and KL-divergence for topic models.
- *DBSCAN.* A widely adopted density-based clustering algorithm [3]. We tune ϵ separately for VSM and the topic models and finally set $\epsilon = 0.5$ for VSM and $\epsilon = 0.7$ for topic models. $minPts$ is set to 1.
- *Direct.* We utilize the topic models learned on the full Twitter corpus to perform "hard clustering" of tweets. Formally, $cluster(d) = \arg\max_k \theta_d^k$. In this way, we obtain C clusters, where C is less or equal to the number of latent topics K.

Results. Figures 6(a) and 6(b) present the overall clustering results. In the figures, "KM" and "DS" refers to K-means and DBSCAN, respectively, and "T-LDA" denotes Twitter-LDA. Due to the large number of obtained results for each topic model, we only present the best result for each model.

Starting with the baseline VSM representation, we observe a relatively high clustering quality when using both K-means and DBSCAN. Importantly, VSM outperforms Twitter-LDA using K-means. Using DBSCAN, VSM outperforms LDA. In fact, this behavior is in agreement with the findings in [15]. Since LDA is based on (potentially sparse) bag-of-words representation of tweets, it fails to produce a significant improvement over VSM.

The tweet representation obtained using MfTM achieves the best overall results using all three clustering algorithms. This shows that the multi-faceted topics from MfTM have better potential to place semantically related tweets into the same clusters. We also note that the performance of the Direct clustering method is better than using the K-means algorithm, while requiring significantly shorter running time. In fact, after training our topic model, Direct only requires constant time to assign a tweet to a cluster.

These experiments demonstrate that the proposed multi-faceted topic model can be effectively applied for clustering short documents, such as tweets.

5. CONCLUSION

In this paper, we study the problem of topic discovery in Twitter. To capture the dynamic and entity-oriented topics in microblogs, we propose a novel Multi-Faceted Topic Model. The model extracts semantically-rich latent topics, including general terms mentioned in the topic, named entities and a temporal distribution. As evidenced by our experimental evaluation, our method demonstrates a high potential to discover more accurate topics for applications such as clustering. Relevant issues for future work include considering the social interactions between users for topic discovery and refining the representation of temporal features of topics.

6. ACKNOWLEDGEMENTS

This work is partially supported by RGC GRF under grant number HKUST 617610.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.
[2] I. Celik, F. Abel, and G.-j. Houben. Learning semantic relationships between entities in twitter. In *ICWE*, 2011.
[3] M. Ester, H. peter Kriegel, J. S, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.
[4] M. Hoffman, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.
[5] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *SOMA Workshop*, 2010.
[6] D. Jiang, K. W.-T. Leung, W. Ng, and H. Li. Beyond click graph: Topic modeling for search engine query log analysis. In *DASFAA*, 2013.
[7] D. Jiang, J. Vosecky, K. W.-T. Leung, and W. Ng. G-wstd: A framework for geographic web search topic discovery. In *CIKM*, 2012.
[8] D. Jiang, J. Vosecky, K. W.-T. Leung, and W. Ng. Panorama: A semantic-aware application search framework. In *EDBT*, 2013.
[9] O. Jin, N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *CIKM*, 2011.
[10] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *ACM AND Workshop*, 2010.
[11] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *KDD*, 2006.
[12] M. Paul and R. Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI*, 2010.
[13] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.
[14] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP*, 2011.
[15] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. Topical clustering of tweets. In *SWSM*, 2010.
[16] J. Vosecky, D. Jiang, and W. Ng. Limosa: A system for geographic user interest analysis in twitter. In *EDBT*, 2013.
[17] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *KDD*, 2006.
[18] W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*. 2011.