4-2011

Comparing Twitter and Traditional Media using Topic Models

Wayne Xin ZHAO
Peking University

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

Jianshu Weng
Peking University

Jing He
Peking University

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

See next page for additional authors

Citation

Zhao, Wayne X., Jiang Jing, Wng Jianshu, He Jing, Lim Ee-Peng, Yan Hongfei and Li Xiaoming. 2011. Comparing Twitter and Traditional Media Using Topic Models. In Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings. Berlin: Springer Verlag.

Published in Advances in Information Retrieval, 2011.
doi:10.1007/978-3-642-20161-5_34
The final publication is available in link.springer.com

Comparing Twitter and Traditional Media using Topic Models

Wayne Xin Zhao[1], Jing Jiang[2], Jianshu Weng[2], Jing He[1], Ee-Peng Lim[2], Hongfei Yan[1] and Xiaoming Li[1]

Peking University, China[1]
Singapore Management University, Singapore[2]

Abstract. Twitter as a new form of social media can potentially contain much useful information, but content analysis on Twitter has not been well studied. In particular, it is not clear whether as an information source Twitter can be simply regarded as a faster news feed that covers mostly the same information as traditional news media. In This paper we empirically compare the content of Twitter with a traditional news medium, New York Times, using unsupervised topic modeling. We use a Twitter-LDA model to discover topics from a representative sample of the entire Twitter. We then use text mining techniques to compare these Twitter topics with topics from New York Times, taking into consideration topic categories and types. We also study the relation between the proportions of opinionated tweets and retweets and topic categories and types. Our comparisons show interesting and useful findings for downstream IR or DM applications.

Keywords: Twitter, microblogging, topic modeling

1 Introduction

Over the past few years, Twitter, a microblogging service, has become an increasingly popular platform for Web users to communicate with each other. Because tweets are compact and fast, Twitter has become widely used to spread and share breaking news, personal updates and spontaneous ideas. The popularity of this new form of social media has also started to attract the attention of researchers. Several recent studies examined Twitter from different perspectives, including the topological characteristics of Twitter [1], tweets as social sensors of real-time events [2], the forecast of box-office revenues for movies [3], etc. However, the explorations are still in an early stage and our understanding of Twitter, especially its large textual content, still remains limited.

Due to the nature of microblogging, the large amount of text in Twitter may presumably contain useful information that can hardly be found in traditional information sources. To make use of Twitter's textual content for information retrieval tasks such as search and recommendation, one of the first questions one may ask is what kind of special or unique information is contained in Twitter. As

Twitter is often used to spread breaking news, a particularly important question is how the information contained in Twitter differs from what one can obtain from other more traditional media such as newspapers. Knowing this difference could enable us to better define retrieval tasks and design retrieval models on Twitter and in general microblogs.

To the best of our knowledge, very few studies have been devoted to content analysis of Twitter, and none has carried out deep content comparison of Twitter with traditional news media. In this work we perform content analysis through topic modeling on a representative sample of Twitter within a three-month time span, and we empirically compare the content of Twitter based on the discovered topics with that of news articles from a traditional news agency, namely, New York Times, within the same time span. Specifically we try to answer the following research questions:

- Does Twitter cover similar categories and types of topics as traditional news media? Do the distributions of topic categories and types differ in Twitter and in traditional news media?
- Are there specific topics covered in Twitter but rarely covered in traditional news media and vise versa? If so, are there common characteristics of these specific topics?
- Do certain categories and types of topics attract more opinions in Twitter?
- Do certain categories and types of topics trigger more information spread in Twitter?

Some of our major findings are the following: (1) Twitter and traditional news media cover a similar range of topic categories, but the distributions of different topic categories and types differ between Twitter and traditional news media. (2) As expected, Twitter users tweet more on personal life and pop culture than world events. (3) Twitter covers more celebrities and brands that may not be covered in traditional media. (4) Although Twitter users tweet less on world events, they do actively *retweet* (forward) world event topics, which helps spread important news.

These findings can potential benefit many Web information retrieval applications. For example, for Web information retrieval and recommendation, our findings suggest that Twitter is a valuable source for entertainment and lifestyle topics such as celebrities and brands to complement traditional information sources. Retweets can also be used to indicate trendy topics among Web users to help search engines refine their results.

2 Data Preparation

We use a sample of the Edinburgh Twitter Corpus [4] as our Twitter dataset. The original corpus was collected through Twitter's streaming API and is thus a representative sample of the entire Twitter stream. It covers a time span from November 11, 2009 to February 1, 2010.

In order to obtain a parallel news corpus that represents the traditional news media, we chose New York Times (NYT) as our source of news articles. We

Table 1. Some statistics of the Twitter and the NYT data sets after preprocessing.

Collection	Docs	Users	Words	Vocabulary
Twitter	1,225,851	4,916	8,152,138	21,448
NYT	11,924	–	4,274,404	26,994

crawled news articles dating from November 11, 2009 until February 1, 2010 through NYT's search page[1].

For both the Twitter and the NYT collections, we first removed all the stop words. We then removed words with a document frequency less than 10 and words that occurred in more than 70% of the tweets (news articles) in the Twitter (NYT) collection. For Twitter data, we further removed tweets with fewer than three words and all the users with fewer than 8 tweets . Some statistics of the two datasets after preprocessing are summarized in Table 1.

3 Topic Discovery and Classification

To compare the content of Twitter and New York Times, we first introduce three major concepts used in this paper.

Definition 1. *A* **topic** *is a subject discussed in one or more documents. Examples of topics include news events such as "the Haiti earthquake," entities such as "Michael Jackson" and long-standing subjects such as "global warming." Each topic is assumed to be represented by a multinomial distribution of words.*

Definition 2. *A* **topic category** *groups topics belonging to a common subject area together. We adopt the topic categories defined in New York Times*[2] *with some modifications. See Figure 3 for the full set of topic categories we use.*

Definition 3. *A* **topic type** *characterizes the nature of a topic. After examining some topics from both Twitter and New York Times, we define three topic types, namely,* event-oriented *topics,* entity-oriented *topics and* long-standing *topics.*

Note that topic categories and topic types are two orthogonal concepts. We assume that each topic can be assigned to a topic category and has a topic type. We use fully automatic methods to discover topics from each data collection first. We then use semi-automatic methods to assign the topics to the predefined topic categories as well as to remove noisy background topics. Finally we manually label the topics with topic types.

3.1 Topic Discovery

New York Times To discover topics from NYT, we choose to directly apply Latent Dirichlet Allocation (LDA) [5]. Our experiments show that we can obtain meaningful topics from the NYT data set using standard LDA. We set the

[1] http://query.nytimes.com/search/
[2] As of July 5, 2010.

number of topics to 100 and ran 1000 iterations of Gibbs sampling using the GibbsLDA++ toolkit[3]. We use $\mathcal{T}_{\text{nyt}}$ to denote the set of topics we obtained from NYT.

Twitter Standard LDA may not work well with Twitter because tweets are short. To overcome this difficulty, some previous studies proposed to aggregate all the tweets of a user as a single document [6, 7]. In fact this treatment can be regarded as an application of the author-topic model [8] to tweets, where each document (tweet) has a single author. However, this treatment does not exploit the following important observation: A single tweet is usually about a single topic. We therefore propose a different Twitter-LDA model.

Formally, we assume that there are T topics in Twitter, each represented by a word distribution. Let ϕ^t denote the word distribution for topic t and $\phi^{\mathcal{B}}$ the word distribution for background words. Let θ^u denote the topic distribution of user u. Let π denote a Bernoulli distribution that governs the choice between background words and topic words. When writing a tweet, a user first chooses a topic based on her topic distribution. Then she chooses a bag of words one by one based on the chosen topic or the background model. The generation process of tweets is described in Figure 1 and illustrated in Figure 2. Each multinomial distribution is governed by some symmetric Dirichlet distribution. We use Gibbs sampling to perform model inference. Due to the space limit we leave out the derivation details and the sampling formulas.

1. Draw $\phi^{\mathcal{B}} \sim \text{Dir}(\beta), \pi \sim \text{Dir}(\gamma)$
2. For each topic $t = 1, \ldots, T$,
   (a) draw $\phi^t \sim \text{Dir}(\beta)$
3. For each user $u = 1, \ldots, U$,
   (a) draw $\theta^u \sim \text{Dir}(\alpha)$
   (b) for each tweet $s = 1, \ldots, N_u$
      i. draw $z_{u,s} \sim \text{Multi}(\theta^u)$
      ii. for each word $n = 1, \ldots, N_{u,s}$
         A. draw $y_{u,s,n} \sim \text{Multi}(\pi)$
         B. draw $w_{u,s,n} \sim \text{Multi}(\phi^{\mathcal{B}})$ if $y_{u,s,n} = 0$ and $w_{u,s,n} \sim \text{Multi}(\phi^{z_{u,s}})$ if $y_{u,s,n} = 1$

Fig. 1. The generation process of tweets.

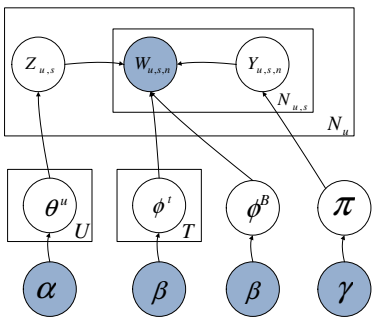

Fig. 2. Plate notation of our Twitter-LDA.

We quantitatively evaluated the effectiveness of our Twitter-LDA model compared with standard LDA model (i.e. treating each tweet as a single document) and the author-topic model (i.e. treating all tweets of the same user as a single document). We set T to 110 (based on preliminary experiments) for these two

[3] http://gibbslda.sourceforge.net/

Table 2. Comparison between Twitter-LDA, author-topic model and standard LDA.

Method	Avg. Score	Agreement between Judges	Cohen's Kappa
Twitter-LDA	**0.675**	65.5%	0.433
Author-Topic	0.539	54.5%	0.323
Standard LDA	0.509	70.9%	0.552

baselines and our Twitter-LDA. We then randomly mixed the 330 topics from the three models. We asked two human judges to assign a score to each topic according to the following guidelines based on the top-10 topic words and took their average as the score for each topic: 1 (meaningful and coherent), 0.5 (containing multiple topics or noisy words), 0 (making no sense). The average scores of topics discovered by each method are shown in Table 2 together with the annotation agreement information. We can see that the Twitter-LDA model clearly outperformed the other two models, giving more meaningful top topic words, indicating that our Twitter-LDA model is a better choice than standard LDA for discovering topics from Twitter.

3.2 Categorizing Topics

New York Times For the NYT dataset, because the articles already have category labels, intuitively, if a topic is associated with many articles in a particular category, the topic is likely to belong to that category. To capture this intuition, we categorize topics by assigning topic t to category q^* where $q^* = \arg\max_q p(q|t) = \arg\max_q p(t|q)p(q)/p(t) = \arg\max_q p(t|q)$, assuming that all categories are equally important. We can estimate the probability of topic t given category q as

$$p(t|q) = \frac{\sum_{d \in \mathcal{D}_{\mathrm{NYT},q}} \tilde{p}(t|d)}{|\mathcal{D}_{\mathrm{NYT},q}|}, \tag{1}$$

where $\tilde{p}(t|d)$ denotes the learned probability of topic t given document d and $\mathcal{D}_{\mathrm{NYT},q}$ denote the subset of documents in the NYT collection that are labeled with category q.

To further remove noisy topics (e.g. topics with incoherent words.) or background topics (e.g. topics consisting mainly of common words such as "called," "made," "added," etc.), we exploit the following observation: Most meaningful topics are related to a single topic category. If a topic is closely related to many categories, it is likely a noisy or background topic. We therefore define a measure called *category entropy (CE)* as follows:

$$\mathrm{CE}(t) = -\sum_{q \in \mathcal{Q}} p(q|t) \log p(q|t). \tag{2}$$

The larger $\mathrm{CE}(t)$ is, the more likely t is a noisy or background topic. We remove topics whose $\mathrm{CE}(t)$ is larger than a threshold (empirically set to 3.41). After removing noisy and background topics, we obtain 83 topics from $\mathcal{T}_{\mathrm{nyt}}$ as the final set of NYT topics we use for our empirical comparison later.

Table 3. Statistics of topics in different types.

Collection	Event-oriented	Entity-oriented	Long-standing
Twitter (81 topics)	7	19	55
NYT (83 topics)	20	9	54

Twitter Unlike NYT documents, tweets do not naturally have category labels. We use the following strategy to categorize Twitter topics. For each Twitter topic we first find the most similar NYT topic. If it is similar enough to one of the NYT topics, we use that NYT topic's category as the Twitter topic's category. Otherwise, we manually assign it to one of the topic categories or remove it if it is a noisy topic. Specifically, to measure the similarity between a Twitter topic t and an NYT topic t', we use JS-divergence between the two word distributions, denoted as p_t and $p_{t'}$:

$$\text{JS-div}(p_t||p_{t'}) = \frac{1}{2}\text{KL-div}(p_t||p_m) + \frac{1}{2}\text{KL-div}(p_{t'}||p_m),$$

where $p_m(w) = \frac{1}{2}p_t(w) + \frac{1}{2}p_{t'}(w)$, and KL-div is the KL-divergence. The JS-divergence has the advantage that it is symmetric. After the semi-automatic topic categorization, we obtain a set of 81 topics from Twitter to be used in later empirical comparison. In the future we will look into automatic methods for cleaning and categorizing Twitter topics.

3.3 Assigning Topic Types

As we described earlier, we have defined three topic types, namely, *event-oriented* topics, *entity-oriented* topics and *long-standing* topics. Because these topic types are not based on semantic relatedness of topics, it is hard to automatically classify the topics into these topic types. We therefore manually classified the Twitter and the NYT topics into the three topic types. Some statistics of the topics in each type are shown in Table 3.

4 Empirical Comparison between Twitter and New York Times

As we have stated, the focus of this study is to compare the content of Twitter with that of New York Times in order to understand the topical differences between Twitter and traditional news media and thus help make better use of Twitter as an information source. In this section we use the discovered topics from the two datasets together with their category and type information to perform an empirical comparison between Twitter and NYT.

4.1 Distribution of Topics

By Topic Categories In traditional news media, while the categories of articles span a wide range from business to leisure, there is certainly an uneven

distribution over these categories. In microblogging sites such as Twitter, where content is generated by ordinary Web users, how does the distribution of different categories of topics differ from traditional news media? To answer this question, we first compute the distributions of different topic categories in NYT and in Twitter respectively in the following way. For NYT, because we have the category labels of news articles, we measure the relative strength of a category simply by the percentage of articles belonging to that category. For Twitter, similarly, we can use the percentage of tweets belonging to each category as a measure of the strength of that category. With the help of the Twitter-LDA model, each tweet has been associated with a Twitter topic, and each Twitter topic is also assigned to a particular category as we have shown in Section 3.2. We also consider an alternative measure using the number of users interested in a topic category to gauge the strength of a category. Only users who have written at least five tweets belonging to that topic category are counted.

We plot the distributions of topic categories in the two datasets in Figure 3, Figure 4 and Figure 5. As we can see from the figures, both Twitter and NYT cover almost all categories. But the relative degrees of presence of different topic categories are quite different between Twitter and NYT. For example, in Twitter, *Family&Life* dominates while this category does not appear in NYT (because it is a new category we added for Twitter topics and therefore no NYT article is originally labeled with this category). *Arts* is commonly strong in both Twitter and NYT. However, *Style* is a strong category in Twitter but not so strong in NYT.

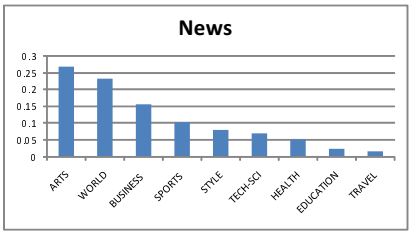

Fig. 3. Distribution of categories in NYT.

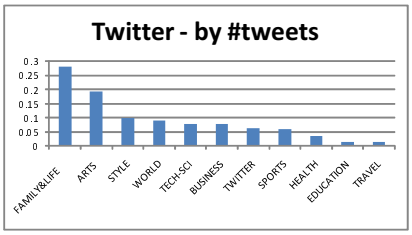

Fig. 4. Distribution of categories (by #tweets) in Twitter.

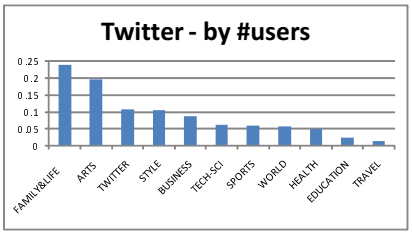

Fig. 5. Distribution of categories (by #users) in Twitter.

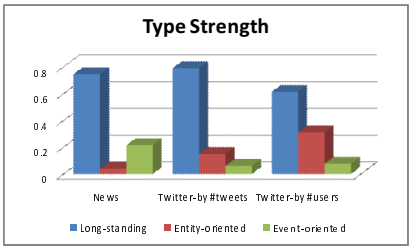

Fig. 6. Distributions of topic types in the two data sets..

By Topic Types Similarly, we can compare the distributions of different topic types in Twitter and in NYT. We show the comparison in Figure 6. An interesting finding is that Twitter clearly has relatively more tweets and users talking about entity-oriented topics than NYT. In contrast, event-oriented topics are not so popular in Twitter although it has a much stronger presence than entity-oriented topics in NYT. We suspect that many entity-oriented topics are about celebrities and brands, and these tend to attract Web users' attention. To verify this, we inspected the entity-oriented topics in Twitter and found that indeed out of the 19 entity-oriented topics in Twitter 10 of them are on celebrities and the other 9 of them are on brands and big companies. Note that long-standing topics are always dominating. It may be surprising to see this for NYT, but it is partly because with LDA model each news article is assumed to have a mixture of topics. So even if a news article is mainly about an event, it may still have some fractions contributing to long-standing topics.

4.2 Breadth of Topic Coverage

Another kind of topic difference is the difference in the breadth of topic coverage. For example, for *Arts*, although both Twitter and NYT have strong presence of this category, we do not know whether they cover roughly the same set of topics. In this section, we first show topics that are covered extensively in Twitter (NYT) but not covered or covered very little in NYT (Twitter). We then try to characterize these topics by ranking topic categories and topic types by their breadth of topic coverage.

Table 4. Topics specific to NYT.

Category	Topics
Arts	book,novel,story,life,writes world,century,history,culture art,museum,exhibition war,history,world,civil,time
Business	cars,car,ford,toyota,vehiclesg media,news,magazine,ads
Edu.	project,money,group,center percent,study,report,rate
Style	french,paris,luxury,swiss,watch
Tech-Sci	space,moon,station,spirit,earth
World	case,charges,prison,trial,court officials,announced,news,week department,agency,federal,law south,north,korea,korean, power

Table 5. Topics specific to Twitter.

Category	Topics
Arts	rob,moon,love,twilight gaga,lady,#nowplaying adam,lambert,fans,kris chirs,brown,song,beyonce download,live,mixtape,music
Business	#ebay,auction,closing #jobs,job,#ukjobs
Family &Life	dog,room,house,cat,door good,night,hope,tonight life,#quote,success,change god,love,lord,heart,jesus smiles,laughs,hugs,kisses
Twitter	tweet,follow,account lmaoo,smh,jus,aint,lmaooo

Twitter-specific and NYT-specific Topics To identify topics present in one dataset but covered very little in the other dataset, we make use of the topic mapping method introduced in Section 3.2. Basically given a topic in Twitter (NYT), we first find its most similar topic in NYT (Twitter) in the same category using the JS-divergence measure. If the divergence measure is above a certain

threshold, meaning that the topic similarity is low, we decide that the topic is not covered in NYT (Twitter). Following Section 3.2 we use a threshold of 0.5 to find Twitter-specific topics and a threshold of 0.504 to find NYT-specific topics. (Both thresholds were set empirically.)

We show some sample specific topics in Table 4 and Table 5. Each topic is shown in one line and represented by a few keywords. First of all, we can see that Twitter-specific topics are concentrated in *Arts* and *Family&Life*. Because we have previously seen that the strength of *Family&Life* is much higher in Twitter than in NYT, it is not surprising to see that this category also has a broader topic coverage than NYT. However, it is interesting to see that although the *Arts* category does not show much difference in terms of relative strength or degree of presence in Twitter and in NYT, its topic coverage is quite different in Twitter and in NYT. In Twitter, there are many specific topics, especially entity-oriented topics such as "Lady Gaga" and "Chris Brown" that are not covered much in NYT. In NYT, there are also certain kinds of topics under *Arts* such as "museum" and "history" that are not covered much in Twitter. In retrospect, if we had separated out a *Pop Culture* category from *Arts*, we might have got different strengths of *Arts* in Twitter and in NYT. On the other hand, many NYT-specific topics are from the category *World*, which is similar to our findings from Section 4.1. It also indicates that news Web sites have broader reports on important events in detail, while due to the length restriction, Twitter tends to report breaking news in brief.

Categories Ranked by Topic Coverage We would like to better characterize the differences of topic coverage of the two data sources in terms of topic categories and types. For topic categories, we would like to see which categories have relative smaller topic coverage in NYT compared with Twitter, and vice versa. To do so, we define the following *topic coverage divergence* (TC-div) measure, which measures the divergence of the topic coverage of one category in Twitter (NYT) with that in NYT (Twitter).

$$\text{TC-div}_{\text{Twitter}}(q) = \frac{\sum_{t \in \mathcal{T}_{\text{Twitter},q}} \min_{t' \in \mathcal{T}_{\text{NYT},q}} \text{JS-div}(p_t || p_{t'})}{|\mathcal{T}_{\text{Twitter},q}|},$$

$$\text{TC-div}_{\text{NYT}}(q) = \frac{\sum_{t \in \mathcal{T}_{\text{NYT},q}} \min_{t' \in \mathcal{T}_{\text{Twitter},q}} \text{JS-div}(p_t || p_{t'})}{|\mathcal{T}_{\text{NYT},q}|}.$$

Here $\mathcal{T}_{\text{Twitter},q}$ denotes the set of topics in Twitter and belonging to category q.

Based on this measure, we can rank the categories for Twitter and for NYT. Table 6 shows the ranking of categories. If a category is ranked high or has a large TC-div value in Twitter (NYT), it means there are many topics in this category that are covered well in Twitter (NYT) but not well in NYT (Twitter).

Types Ranked by Topic Coverage Similarly, we can also rank the topic types by their topic coverage divergence measures. For both NYT and Twitter, we have the ranking: *Entity-oriented* > *Long-standing* > *Event-oriented*. Event-oriented

type has the smallest TC-div for both news and Twitter while entity-oriented type has the largest TC-div for both news and Twitter. It suggests that Twitter and NYT have more overlap of event-oriented topics but less overlap of entity-oriented topics. Also, event-oriented type has a smaller TC-div in Twitter than in NYT, suggesting that NYT covers event-related content of Twitter well but Twitter does not cover that of NYT quite well.

4.3 Opinions in Twitter

One characteristic of Twitter content compared with traditional news media is arguably the amount and coverage of user opinions expressed in tweets. We further study what categories and types of topics can generate a large number of opinionated tweets. We use a sentiment lexicon of 50 opinionated words[4] to identify opinionated tweets. We roughly estimate the proportions of tweets in each category that are opinionated by the number of tweets in each topic category or topic type that contain at least one of the opinion words. We show the results in Table 7. Interestingly, we can see that while the category *Education* is not a popular topic category in terms its total number of tweets, its proportion of opinionated tweets is ranked high, right after *Family&Life*. Categories such as *Tech-Sci*, *Business* and *World*, whose popularity in Twitter is in the mid-range, have been pushed down to the bottom in terms of their proportions of opinionated tweets. This change of ranking suggests that Twitter users tend to use Twitter to spread news in these categories rather than discuss their own opinions on news in these categories. On the other hand, more life and leisure-related topic categories such as *Style*, *Travel* and *Sports* tend to trigger more personal opinions.

Twitter	NYT
Arts	Education
Family&Life	Style
Business	Art
Travel	Travel
Tech-Sci	World
Health	Business
Education	Health
Style	Tech-Sci
World	Sports
Sports	—

Table 6. Ranking of topic categories based on topic coverage divergence.

Category	Opinion proportion
Family&Life	0.355
Education	0.294
Arts	0.289
Style	0.257
Twitter	0.242
Sports	0.226
Travel	0.198
Health	0.189
Business	0.186
Tech-Sci	0.151
World	0.097

Table 7. Opinion proportions of different categories in Twitter.

Category	Retweet proportion
World	0.359264
Travel	0.22061
Tech-Sci	0.209646
Sports	0.187932
Twitter	0.182681
Style	0.170511
Arts	0.155924
Family&Life	0.141174
Health	0.155875
Business	0.11262
Education	0.082559

Table 8. Retweet proportions of different categories in Twitter.

Similarly, we can do this with topic types. For opinion proportions, *Long-standing* > *Entity-oriented* > *Event-oriented*. As we can see, long-standing top-

[4] We manually went through our Twitter data and selected the top 50 opinion words based on our own judgment.

ics attract more opinionated tweets. It is interesting to see that entity-oriented topics attract relatively more opinions than event-oriented topics. This may be because many event-oriented topics actually also belong to the *World* and *Business* categories, while many entity-oriented topics are related to celebrities and brands, which are more closely related to life and leisure.

4.4 Topic Spread through Retweet

Another special property of Twitter is that it allows people to spread news through *retweet* messages. We further compute the proportions of retweet messages in each topic category and topic type by identifying the pattern `RT: @username`. From Table 8, we can see that the category *World* has the most retweet proportion among all categories. For Retweet proportion of topic types, we get *Entity-oriented* > *Long-standing* > *Event-oriented*. Event-oriented type has the most retweet proportion among all types. This makes sense because many topics in the *World* category also belong to event-oriented topic type, e.g., topics on breaking-news such as "Haiti earthquake." This observation is interesting because although our previous analysis has shown that the strength and breadth of topic coverage of *World* topics in Twitter is low, we do see that Twitter users most actively spread *World* topics than other topics. It shows that retweeting is an important way for dissemination of significant events.

5 Related Work

Recently Twitter has attracted much attention in the research community, e.g. [6, 1]. Our work is quite different from many pioneering studies on Twitter because we try to compare the content differences between Twitter and traditional news media. In terms of topic modeling, our model is based on [8] but samples a single topic for a whole sentence. Recently, [9] applied labeled-LDA to Twitter, but the model relies on hashtags in Twitter, which may not include all topics. [7] conducted an empirical study of different strategies to aggregate tweets based on existing models. Our proposed Twitter-LDA differs from the models studied in [7] in that we model one tweet with one topic label, which is similar to [10, 11] but for different applications. Another related area is comparison of text corpora [12–14]. The nature of Twitter makes our work more difficult than previous studies because tweets are short messages and different from traditional documents. In addition, no previous work has compared topics in different views, i.e. topics of differen categories and topics of different types. A most recent piece of work [15] tries to explore search behavior on the popular microblogging site Twitter, which also has a different focus than ours.

6 Conclusions

In this paper we empirically compared the content of Twitter with a typical traditional news medium, New York Times, focusing on the differences between

these two. We developed a new Twitter-LDA model that is designed for short tweets and showed its effectiveness compared with existing models. We introduced the concepts of topic categories and topic types to facilitate our analysis of the topical differences between Twitter and traditional news media. Our empirical comparison confirmed some previous observations and also revealed some new findings. In particular, we find that Twitter can be a good source of entity-oriented topics that have low coverage in traditional news media. And although Twitter users show relatively low interests in world news, they actively help spread news of important world events.

In the future, we will study how to summarize and visualize Twitter content in a systematic way. Our method of associating tweets with different categories and types may also help visualization of Twitter content.

Acknowledgement This work was done during Xin Zhao's visit to the Singapore Management University. Xin Zhao, Hongfei Yan and Xiaoming Li are are partially supported by NSFC under the grant No. 70903008, 60933004, 61073082 and 61050009, CNGI grant No. 2008-122 and Grant No. SKLSDE-2010KF-03, Beihang University.

References

1. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th WWW. (2010)
2. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th WWW. (2010)
3. Asur, S., Huberman, B.A.: Predicting the future with social media. WI-IAT (2010)
4. Petrović, S., Osborne, M., Lavrenko, V.: The Edinburgh Twitter corpus. In: Proceedings of the NAACL HLT 2010 Workshop. (2010)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR (2003)
6. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM WSDM. (2010)
7. Hong, L., Davison, B.D.: Empirical study of topic modeling in Twitter. In: Proceedings of the SIGKDD Workshop on SMA. (2010)
8. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: SIGKDD. (2004)
9. Ramage, D., Dumais, S., Liebling, D.: Characterizing micorblogs with topic models. In: Proceedings of AAAI on Weblogs and Social Media. (2010)
10. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proceeding of the 17th WWW. (2008)
11. Li, P., Jiang, J., Wang, Y.: Generating templates of entity summaries with an entity-aspect model and pattern mining. In: Proceedings of the 48th ACL. (2010)
12. Zhai, C., Velivelli, A., Yu, B.: A cross-collection mixture model for comparative text mining. In: Proceedings of the tenth ACM SIGKDD. (2004)
13. Paul, M., Girju, R.: Cross-cultural analysis of blogs and forums with mixed-collection topic models. In: Proceedings of the 2009 EMNLP. (2009)
14. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD. (2009)
15. Teevan, J., Ramage, D., Morris, M.: #Twittersearch: A comparison of microblog search and web search. In: Proceedings of the fourth ACM WSDM. (2011)