Singapore Management University
School of Information Systems
Technical Paper Series

An Empirical Comparison of Topics in Twitter and Traditional Media

Xin Zhao
School of Electrical Engineering and Computer Science
Peking University
Beijing, China
batmanfly@gmail.com

Jing Jiang
School of Information Systems
Singapore Management University
Singapore
jingjiang@smu.edu.sg

January, 2011

An Empirical Comparison of Topics in Twitter and Traditional Media

Wayne Xin Zhao[1] and Jing Jiang[2]
Peking University, China[1]
Singapore Management University, Singapore[2]

batmanfly@gmail.com, jingjiang@smu.edu.sg

January 20, 2011

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 of topic categories and types. We find that although Twitter and New York Times cover similar categories and types of topics, the distributions of topic categories and types are quite different. Furthermore, there are Twitter-specific topics and NYT-specific topics, and they tend to belong to certain topic categories and types. We also study the relation between the proportions of opinionated tweets and retweets and topic categories and types, and find some interesting dependence. To the best of our knowledge, ours is the first comprehensive empirical comparison between Twitter and traditional news media.

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. Twitter allows users to broadcast short posts (called "tweets") online in real time, often through mobile phones. Each tweet is limited to a maximum length of 140 characters. "Followers" of a Twitter user (called a Twitterer) are constantly notified of that Twitterer's tweets, while everyone else can also access these tweets online. 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 [4], tweets as social sensors of real-time events [11], the sentiment prediction power of Twitter [15], etc. However, the explorations are still in an early stage and our understanding of Twitter still remains limited.

To understand and subsequently make use of Twitter (or microblogging in general), one of the first questions one may ask is what kind of useful information is contained in Twitter's large body of content. Here we do not consider Twitter's utility for individual users as a networking tool among friends. We focus mainly on Twitter in its entirety as a new form of information source that can be used for search and recommendation, for example. As Twitter is often used to spread breaking news, a particularly important question to ask is how the information contained in Twitter differs from what one can obtain from other more traditional media such as newspapers.

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. [4] examined the "trending topics" published by Twitter, which represent the top ten most popular keywords at any given time, and compared them with Google Trend and CNN Headline News. However, these keyword-based trending topics only capture a small subset of topics in Twitter, and this topic comparison was not the focus of their study. [10, 3] performed unsupervised topic modeling on Twitter, which covers almost all topics in Twitter, but there was no comparison with traditional news media.

In this work we perform content analysis through topic modeling on a representative sample of the entire 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 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?

Because Twitter presumably reflects the real interests of today's Web users, many of whom are also online consumers, answers to the questions above can help uncover these user interests and subsequently facilitate many applications that may benefit from such knowledge. For example, by identifying popular topics in Twitter that are not extensively reported or not covered at all by traditional news media, we can filter out redundant information and discover novel information contained in Twitter. By identifying the categories and types of topics that tend to attract more Web users' attention, advertisers may design better strategies for placing advertisements not only online but also in traditional media, which may not have an effective feedback loop.

We start with topic discovery from Twitter and from New York Times, a typical traditional news medium, employing LDA-based topic models. For Twitter, we develop a new Twitter-LDA model which handles short tweets better than standard LDA. We then classify the discovered topics into a set of predefined topic categories and types. We conduct empirical comparison of the topics in Twitter and in New York Times, using topic categories and types to characterize the differences.

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) Relatively speaking Twitter users tweet less on world events than on personal life and pop culture. (3) Twitter covers more entity-oriented topics on celebrities and brands which may not be covered in traditional media. (4) Although Twitter users seem to be less interested in world events, they do actively retweet world event topics, which helps spread important news.

The rest of the paper is organized as follows. We first describe our data preparation process in Section 2. We then present the algorithms to perform topic modeling and classification in Section 3. In Section 4 we empirically compare Twitter with New York Times using the topics discovered from them. We discuss related work in Section 5 and conclude in Section 6.

2 Data Preparation

Twitter Data

The Twitter data we use comes from the Edinburgh Twitter Corpus [9]. The original corpus was collected from November 11, 2009 until February 1, 2010 and contains 9 million users and 96 million tweets. The data was collected through Twitter's streaming API and is thus a representative sample of the content in the entire Twitter during that period. Because the original corpus is huge and we have limited computing resources, we further pruned the corpus. We tried to minimize the bias introduced while still keeping the most representative tweets in the following pruning steps. We first selected the top 20,000 users with the most tweets. We then removed tweets containing non-English letters and tweets no longer than 10 characters. Next we tried to filter out those Twitter bots from real users because these bots can automatically generate a huge number of tweets that dominate the corpus but do not represent the interests of human

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

users. It was not easy to find a simple yet accurate rule to separate these bots, and in the end we chose to remove users who had not relied to any tweet, i.e. no retweet (RT) messages. Finally, we manually removed all non-English users. By default, we treat each tweet as a single document.

New York Times Data

In order to obtain a parallel news corpus, we chose New York Times (NYT) as our source of news articles. We crawled news articles through NYT's search page[1]. We ignored pages that could not be downloaded or processed, and obtained 11,924 articles spanning also from November 11, 2009 until February 1, 2010. We treat each article as a single document.

Preprocessing

For both the Twitter and the NYT collections, we first removed all the stop words. We then removed both low-frequency and high-frequency words as follows. We removed words with a document frequency less than 10, and we removed words that occurred in more than 70% of the tweets (news articles) in the Twitter (NYT) collection. Some statistics of the two data sets after preprocessing are summarized in Table 1.

3 Topic Discovery and Classification

To compare the content of Twitter and New York Times, we first identify the topics covered in each data set.

Definition 1 *A* topic *is a subject discussed in one or more documents. Examples of topics include news events such as "the Haiti earthquake" and "the Iranian election," entities such as "Michael Jackson" and "the Los Angeles Lakers," and long-standing subjects such as "music" and "global warming."*

Because our data sets are large, it is only feasible to use fully unsupervised or weakly supervised methods to automatically discover topics. Here we adopt LDA-based topic modeling to perform topic discovery. Each topic is assumed to be represented by a multinomial distribution of words.

[1] http://query.nytimes.com/search/

Besides using individual topics to compare Twitter and NYT, we also introduce the following two concepts in order to better characterize the content similarities and differences between Twitter and NYT.

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 such as merging* Arts*,* Theater*,* Book *and* Movie *together as* Arts*, merging* U.S. *and* World *as* World*. See Table 3 for the full set of topic categories.*

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 belongs to a topic type.

In the rest of this section, we present the details of our topic modeling and topic classification methods. 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 from NYT

To discover topics from NYT, we choose to directly apply Latent Dirichlet Allocation (LDA) [1] because news articles from NYT are generally long and topically coherent and thus suitable for standard LDA. Our experiments also showed that we could obtain meaningful topics from the NYT data set using standard LDA. We set the 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.

3.2 Topic Discovery from Twitter

Although we could also apply LDA to discover topics from tweets by treating each tweet as a single document, this direct application would most likely not work well because tweets are very short, often containing only a single sentence. To overcome this difficulty, some previous studies proposed to aggregate all the tweets of a user as a single document [17, 3]. In fact this treatment can be regarded as an application of the author-topic model [12] to tweets, where each document (tweet) has a single author. However, the discovered topics are sometimes confusing because the aggregated tweets of a single user may have a diverse range of topics. On the other hand, this model does not exploit the

[2] We last crawled the NYT data on July 5, 2010 and we found that the categories at the Web site had since been changed.

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

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

Figure 1: The generation process of tweets.

following important observation: A single tweet is usually about a single topic. This assumption makes uses of the length restriction in Twitter. We therefore propose a different Twitter-LDA model.

3.2.1 Model Description

Our model is based on the following assumptions. There are T topics in Twitter, each represented by a word distribution. Each user has her topic interests and therefore a distribution over the T topics. When a user wants to write a tweet, she first chooses a topic based on her topic distribution. Then she chooses a bag of words one by one based on the chosen topic. However, not all words in a tweet are closely related to the topic of that tweet; some are background words commonly used in tweets on different topics. Therefore, for each word in a tweet, the user first decides whether it is a background word or a topic word and then chooses the word from its respective word distribution.

Formally, 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. 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.

3.2.2 Model Inference

We use Gibbs sampling to perform model inference. Due to the space limit we leave out the derivation details and only show the sampling formulas below. We use $\boldsymbol{w}$ to denote all the words we observe in the collection, and $\boldsymbol{y}$ and $\boldsymbol{z}$ to denote all the hidden variables.

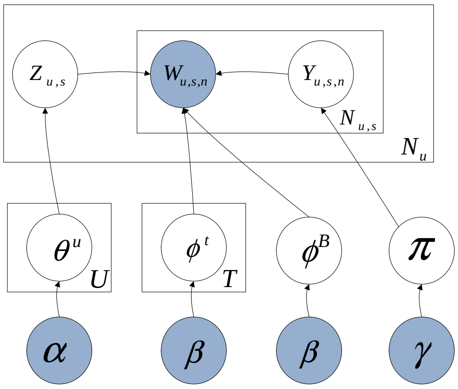

Figure 2: Plate notation of our Twitter-LDA.

First, given the assignment of all other hidden variables, to sample a value for $z_{u,s}$, we use the following formula:

$$P(z_{u,s} = t|\boldsymbol{z}_{\neg(u,s)}, \boldsymbol{y}, \boldsymbol{w}) \propto \frac{c^u_{(t)} + \alpha}{c^u_{(\cdot)} + T\alpha} \times \left(\frac{\Gamma\left(c^t_{(\cdot)} + V\beta\right)}{\Gamma\left(c^t_{(\cdot)} + n^t_{(\cdot)} + V\beta\right)} \cdot \prod_{v=1}^{V} \frac{\Gamma\left(c^t_{(v)} + n^t_{(v)} + \beta\right)}{\Gamma\left(c^t_{(v)} + \beta\right)} \right).$$

Here $c^u_{(t)}$ is the number of tweets assigned to topic t for user u, and $c^u_{(\cdot)}$ is the number of tweets that user u has posted. $c^t_{(v)}$ is the number of times word v is assigned to topic t, and $c^t_{(\cdot)}$ is the number of times any word is assigned to topic t. All the counts above exclude the current tweet s of user u. $n^t_{(v)}$ is the number of times word v is assigned as a topic word to topic t in tweet s of user u, and $n^t_{(\cdot)}$ is the number of times any word is assigned as a topic word to topic t in tweet s of user u.

Then, given $z_{u,s} = t$, to jointly indicator values for $y_{u,s,n}$, we have

$$P(y_{u,s,n} = 0|\boldsymbol{z}, \boldsymbol{y}_{\neg(u,s,n)}, \boldsymbol{w}) \propto \frac{c^{\mathcal{B}}_{(v)} + \beta_2}{c^{\mathcal{B}}_{(\cdot)} + V\beta} \cdot \frac{c_{(0)} + \gamma}{c_{(\cdot)} + 2\gamma},$$

$$P(y_{u,s,n} = 1|\boldsymbol{z}, \boldsymbol{y}_{\neg(u,s,n)}, \boldsymbol{w}) \propto \frac{c^t_{(v)} + \beta}{c^t_{(\cdot)} + V\beta} \cdot \frac{c_{(1)} + \gamma}{c_{(\cdot)} + 2\gamma}.$$

$c^t_{(v)}$ is the number of times word v is assigned to topic t. $c^{\mathcal{B}}_{(v)}$ is the number of times word v is assigned to the background model. $c_{(0)}$ is the number of words assigned as background words and $c_{(1)}$ is the number of words assigned as topic words.

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

Method	Average Score	Agreement (#matched/#topics)	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

3.2.3 Model Evaluation

When applying the Twitter-LDA model, we set $\alpha = T/50$ and $\beta = 0.01$, which are the default settings suggested in [2]. For γ, after trying several values we set it to 20. As for the number of topics, we also tried a range of values and we found that a number between 100 and 120 was a good choice for our data set. We thus chose $T = 110$. We ran 400 iterations of Gibbs sampling of our Twitter-LDA model, and then collected 10 samples with a lag of 10 iterations between two samples.

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), using manually annotated topics. First, we applied our Twitter-LDA, standard LDA and the author-topic model to the Twitter data set. For standard LDA and the author-topic model, we also set T to 110 and we ran 1000 iterations of Gibbs sampling. We then randomly mixed the 330 topics from the three models and presented the top ten words of each topic to two human judges. We asked the human judges to assign a score to each topic according to the following guidelines: If the top ten words are meaningful and coherent, a score of 1 is assigned to the topic; If the top ten words suggest multiple topics or if there are noisy words, a score of 0.5 is assigned; If it is impossible to make any sense out of the top ten words, a score of 0 is assigned.

In Table 2, we show the average score of the topics discovered by each model, the percentage of agreed annotations between the two human judges and Cohen's Kappa. We can see that the Twitter-LDA model clearly outperformed the other two models, giving more meaningful top topic words. This comparison shows that our Twitter-LDA model is a good choice for discovering topics from Twitter.

3.3 Categorizing NYT Topics

Given our categories taken from NYT with slight modifications, now we describe how we use a semi-automatic method to categorize the NYT topics discovered by LDA as well as to remove noisy background topics.

For the NYT data set, because the articles already have category labels, intuitively, if a topic is associated with many articles in a particular category, then the topic is likely to belong to that category. To capture this intuition, we categorize topics as follows. First, after unsupervised LDA learning we can obtain a topic distribution for each document. Let $\tilde{p}(t|d)$ denote the learned probability of topic t given document d. For a topic category q, let $\mathcal{D}_{\mathrm{NYT},q}$ denote the subset of documents in the NYT collection that are labeled with category q. Then we can define the probability of topic t given category q as

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

Note that $p(t|q)$ is a valid probability measure because we have $\sum_{t \in \mathcal{T}_{\mathrm{NYT}}} p(t|q) = 1$. If we assume that all categories are equally important, we have

$$p(q|t) = \frac{p(t|q)p(q)}{\sum_{q' \in \mathcal{Q}} p(t|q')p(q')} = \frac{p(t|q)}{\sum_{q' \in \mathcal{Q}} p(t|q')},$$

where $\mathcal{Q}$ is the set of all topic categories. We can then assign topic t to category q^* where

$$q^* \quad = \quad \arg\max_{q} p(q|t).$$

However, some of the topics generated by LDA are noisy background topics and we want to remove these topics. We exploit the following observation: Most meaningful topics are related with a single topic category. If a topic is closely related with many categories, it is likely a background topic. We therefore define a measure called *category entropy (CE)* as follows:

$$\mathrm{CE}(t) \quad = \quad -\sum_{q \in \mathcal{Q}} p(q|t) \log p(q|t). \qquad (1)$$

The larger $\mathrm{CE}(t)$ is, the more likely t is a background topic.

We remove topics whose $\mathrm{CE}(t)$ is larger than a threshold. We empirically set this threshold to 3.41. After removal of background topics, we obtained 83 topics from $\mathcal{T}_{\mathrm{nyt}}$ as the final set of NYT topics we use for our empirical comparison later.

3.4 Categorizing Twitter Topics

Unlike NYT documents, tweets do not naturally have category labels, so we cannot use the same automatic method as for NYT to map Twitter topics to the categories. We take the following approach. For each Twitter topic we first find the most similar NYT topic. If the similarity is too small based on a threshold we empirically set, then we assume that this Twitter topic does not have a corresponding NYT topic and we manually assign it to one of the topic categories or remove it if it is a noisy topic; otherwise, if it is similar enough to one of the NYT topics, we use that NYT topic's category as the Twitter topic's category.

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. The smaller the JS-div is, the more similar two topics are.

After the semi-automatic topic categorization, we obtained a set of 81 topics from Twitter to be used in our empirical comparison later.

We give some sample topics in Twitter and in NYT together with their categories in Table 3.

3.5 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 4.

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 a representative traditional news medium, namely, 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.

Now that we have discovered the topics from our Twitter data set and NYT data set using topic modeling and have classified the topics into categories and types, in this section we use these topics together with their category and type information to perform an empirical comparison between Twitter and NYT.

4.1 Distribution of Topics

Probably what is most interesting is to compare the most popular topics in each data set. For NYT, we observe that there are two special sections of news articles, namely, *FrontPage* and *Opinion* articles. *FrontPage* is the collection of news in the front pages of New York Times, which naturally contains the most important news, while *Opinion* contains the collection of opinion articles, which are likely on news that attracts much attention from the readers. So we use these two sections to extract popular topics from NYT.

Table 3: Sample topics in each category for NYT and Twitter.

Categories	Topics from NYT	Topics from Twitter
Arts	dance,ballet,theater,dancers,arts	rob, moon, love, twilight,edward
	world, century,history,social,culture	herlock, holmes, #holmes
	art, museum, exhibition, paintings	gaga, lady, #nowplaying
Business	percent,market,prices,rose,quarter,fell	free, products, store,online
	bank,financial,government,debt,money, loans	money, making, business, online
	jobs,economic,economy,unemployment	#ebay, auction, closing
Education	university, students, college, education	school, class, english, math
	school,high,teachers,class,children	
	project,money,group,development,center	
Health	health, care, bill, senate, house,insurance	health, flu, swine, #h1n1, #swineflu
	human, research,cells,brain,disease,cancer	care, anti, aging, skin
	research, data, head, brain, concussions	bad, sick, hurts, head, cough, pain
Sports	game,season,football,play,team	lakers, kobe, nba, play
	world, cup, team,soccer,Africa	cup, world, united, fifa, africa
	open, Australia, tennis,final, match	tiger, woods, golf, play
Style	fashion,designer,look,made,clothes	hair, black, cut,red, skin
	french, paris, luxury, swiss,watch	wear, jeans, shoes, black
	wine,bar,beer,drink,tea,bottles	party, night,club, ladies
Tech-Sci	apple,technology,iphone,mobile,computer	iphone, #iphone, apple, app
	google, internet, web, online,search, site	video, youtube, feature, autoshare
	space, moon, station, spirit,earth	#gamer, #gaming, ps3, xbox
Travel	city, hotel,street, building,house,room	fishing, car, bicycle, cars
	island, sea, coast,beach,boat,ocean,miles	city,south, coast, east, west
	flight, travel,air, airport,passengers	travel, #travel, #tips, hotel
World	haiti,earthquake,relief,help,january	haiti, #haiti, earthquake, donate
	nuclear,iran,weapons,security,sanctions	#iranelection, #iran, iran, arrested
	Afghanistan,American,pakistan,taliban	#news, #tcot, police, pakistan
Family & Life		mum, dad, home, family
		night, good, sleep, dreams, sweet
		love, cry,feel, smile
Twitter		twitter, tweet, follow, account
		follow,followers, twitter
		lmaoo, smh, jus, aint, lmaooo

Table 4: 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

Let $\mathcal{D}_{\text{FrontPage}}$ denote the set of FrontPage articles. We measure the popularity of a topic in the FrontPage section by the following formula, which represents

Table 5: The top 5 most popular topics from the two data sets.

Source	**Top 5 Most Popular Topics**
FrontPage (NYT)	security,qaeda,yemen,attacks, american
	Afghanistan,American,pakistan,taliban
	obama,president,white,house,clinton,policy
	health, care, bill, senate, house,insurance
	party,political,election,republican, voters
Opinion (NYT)	health, care, bill, senate, house,insurance
	obama,president,white,house,clinton,policy
	department,agency,federal,law,rules,commision
	party,political,election,republican, voters
	jobs,economic, spending, unemployment,budget
Twitter	ppl, love, life, feel
	twitter, tweet, follow, account
	good, time, night, home, mum
	sleep, bed, night, asleep, tired
	song , love , music , listening

on average the probability of an article talking about topic t:

$$s_{\text{FrontPage}}(t) = \frac{\sum_{d \in \mathcal{D}_{\text{FrontPage}}} \tilde{p}(t|d)}{|\mathcal{D}_{\text{FrontPage}}|},$$

where $\tilde{p}(t|d)$ is obtained from LDA. We can compute the popularity score $s_{\text{Opinion}}(t)$ in a similar way.

For Twitter, because our Twitter-LDA model assigns a single topic to each tweet, we simply use the number of tweets assigned to a topic to measure its popularity.

We show the top five most popular topics from the *FrontPage* and *Opinion* sections of NYT as well as from Twitter in Table 5. We can see that none of the most popular Twitter topics is about any world news; they are all long-standing topics related to everyday life. In contrast, the popular *FrontPage* NYT topics are all headline news, which is not surprising. The popular *Opinion* NYT topics are either related to news events (topics 1, 2 and 4) or some politics or business-related long-standing topics (topics 3 and 5). This comparison clearly shows that the most popular topics in Twitter are very different from those in traditional news media; popular topics in Twitter tend to be about everyday life.

To better characterize the comparison between Twitter and NYT in their topic distributions, we use the topic category and type information to further examine the differences.

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. Tweets are not naturally labeled with topic categories, but 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 3.4. Thus it is not hard to obtain the percentage of tweets in each topic category.

However, this number-of-tweets-based measure may be biased by tweets generated by more active Twitter users. In traditional news media where article publication is centrally controlled or moderated by a chief-editor, there is not such a problem of a small number of authors dominating the coverage of topics. In Twitter, however, where publication is free from any central control, number of tweets may not be a good indicator of the overall user interests. We therefore also consider an alternative measure using the number of users interested in a topic category to gauge the strength of a category. To do this, for a given topic category, we count all users who have written at least five tweets belonging to that topic category.

We plot the distributions of topic categories in the two data sets 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.

Comparing Figure 4 with Figure 5, we can also see that the category *World* has a significant drop from ranking by #tweets to ranking by #users. It indicates that only a moderate proportion of Twitter users are interested in *World* topics. It is consistent with the findings in [4], where the authors found that although topics “apple” and “iran-election” had similar numbers of tweets, the number of users participating in conversations on “apple” was five times larger than that of “iran-election.” We can also observe that Twitter has a more balanced distribution of different categories than in news excluding the top two categories.

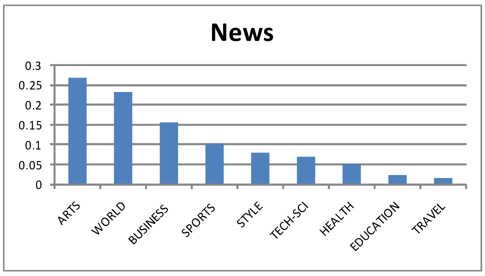

Figure 3: Distribution of categories in NYT.

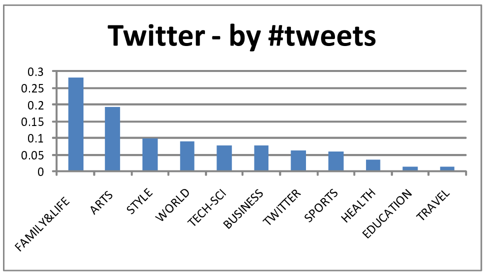

Figure 4: Distribution of categories (by #tweets) in Twitter.

By Topic Types

We would also like to compare the distributions of different topic types in Twitter and in NYT. Recall that we have three topic types, namely event-oriented topics, entity-oriented topics and long-standing topics. Using measures similar to the ones defined in the previous section, we can also obtain these distributions and plot them out. 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

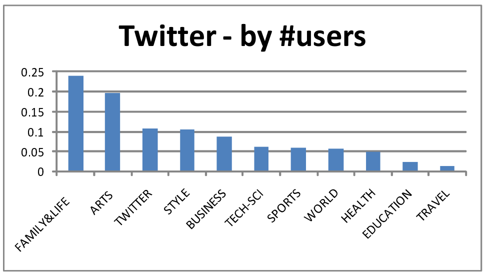

Figure 5: Distribution of categories (by #users) in Twitter.

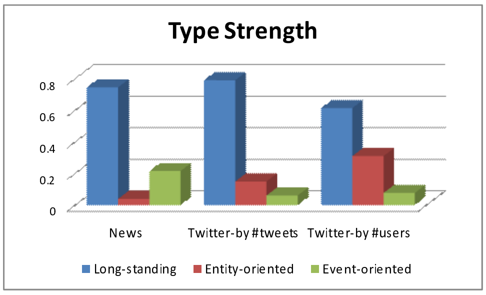

Figure 6: Distributions of topic types in the two data sets.

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 the dominating one. 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 has some fractions contributing to long-standing topics.

4.2 Breadth of Topic Coverage

In the previous section we compared the relative degrees of presence or the strengths of different topic categories and topic types in Twitter and in NYT. 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 such as books and movies. 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.

Twitter-specific and NYT-specific Topics

To identify topics present in one data set but covered very little in the other data set, we make use of the topic mapping method introduced in Section 3.4. 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).

Table 6: Topics specific to NYT.

Categories	**Specific Topics from news**
ARTS	book, novel,story, life,writes,writer world, century,history,social,culture art, museum, exhibition, paintings television, nbc,time, network, cable, fox british ,london,england ,royal war,history, world,civil, time publishers,list,reading,amazon, authors
BUSINESS	cars, car, ford, toyota,vehicles,diriving media, news, magazine, advertising, ads
EDU.	project,money,group,development,center percent, study, report, rate, average
STYLE	french, paris, luxury, swiss,watch
TECH-SCI	space, moon, station, spirit,earth
WORLD	case,charges,prison,trial,court,justice officials,announced,news,week, statement department,agency,federal,law,rules court, case, judge, supreme, lawyers,justice south,north,korea,korean, power european, europe, union, russian, germany

Following Section 3.4 we use a threshold of 0.5 to find Twitter-specific topics and a threshold of 0.504 (set empirically) to find NYT-specific topics.

We show these specific topics in Table 6 and Table 7. First of all, as we can see from the tables, 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 details, while due to the length restriction, Twitter tends to report breaking news in brief.

Table 7: Topics specific to Twitter.

Categories	Specific topics from twitter
ARTS	rob, moon, love, twilight,edward herlock, holmes, #holmes gaga, lady, #nowplaying adam, lambert, fans, kris chirs, brown, song, beyonce download, live, mixtape, music #nowplaying, #mm, #musicmonday
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,giggles ppl, love, life, feel night, good, sleep, dreams, sweet feel, good, bad, pretty love, cry,feel, smile
TWITTER	twitter, tweet, follow, account follow,followers, twitter lmaoo, smh, jus, aint, lmaooo

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 8 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).

Table 8: Ranking of topic categories based on topic coverage divergence.

Rank	Twitter	NYT
1	Arts	Education
2	Family&Life	Style
3	Business	Art
4	Travel	Travel
5	Tech-Sci	World
6	Health	Business
7	Education	Health
8	Style	Tech-Sci
9	World	Sports
10	Sports	

Table 9: Ranking of topic types based on topic coverage divergence.

Types	**TC-div of NYT**	**TC-div of Twitter**
Entity-oriented	0.489	0.495
Long-standing	0.467	0.483
Event-oriented	0.449	0.410

Types Ranked by Topic Coverage

Similarly, we can also rank the topic types by their topic coverage divergence measures. The rankings are shown in Table 9. As we can see, 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. Traditional news media only contain the opinions of a small number of journalists and possibly some readers, but Twitter allows any user to express her opinions towards anything. Since we have classified topics in Twitter into topic categories and types, it would be interesting to study what categories and types of topics can generate a large number of opinionated tweets.

To do this, we use a sentiment lexicon to identify opinionated tweets. We took a simple approach by manually going through the most frequent words in the Twitter data set and select the top 50 opinion words based on our own judgment. We then count the number of tweets in each topic category or topic type that contain at least one of the opinion words. By doing so, we can obtain a rough estimation of the proportions of tweets in each category that

Table 10: Topic categories ranked by their proportions of opinionated tweets.

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 11: Topic types ranked by their proportions of opinionated tweets.

Type	**Opinion Proportion**
Long-standing	0.284
Entity-oriented	0.219
Event-oriented	0.088

are opinionated.

We show the results in Table 10. 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.

Similarly, we can do this with topic types. We show the results in Table 11. As we can see, long-standing topics 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. These tweets can be easily identified by the pattern `RT: @username`. Different from traditional news media, these retweet messages

Table 12: ReTweet proportions of different types in Twitter.

Type	ReTweet proportion
Event-oriented	0.314
Long-standing	0.162
Entity-oriented	0.144

Table 13: ReTweet 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

strongly indicate the kinds of news that people find interesting or important and thus actively spread.

We therefore also compute the proportions of retweet messages in each topic category and topic type.

From Table 13 and Table 12, we can see that the category *World* has the most retweet proportion among all categories and 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, a new form of social media, has attracted much attention in the research community. [16] defined a network-theoretic model of social awareness stream, which allowed researchers to systematically define and compare different stream aggregations. [17] proposed a novel ranking algorithm for finding topic-sensitive influential twitters. [4] was the first quantitative study on the entire Twittersphere and information diffusion on it. Our work is quite different from these pioneering studies on Twitter because we try to compare the content differences between Twitter and traditional news media.

Our work is based on previous work on topic modeling (e.g. [1, 14, 12]). [14] first proposed to learn rateable aspects by capturing local contexts; [6] implemented this idea by sampling a single topic assignment for a whole sentence. Our model is based on [12] and but also samples a single topic for a whole sentence. Recently there has been some work focusing on extracting topics from Twitter [3, 10]. [10] applied Labeled-LDA to Twitter, but the model relies on hashtags in Twitter, which may not include all topics. [3] conducted an empirical study of different strategies to aggregate tweets based on existing models. The difference between our proposed Twitter-LDA and the models studied in [3] is that we proposed to model one tweet with one topic label.

Another related research field is comparison of text corpora. [18] extended basic PLSA model to model collection-independent topics and collection-specific topics. [8] further improved [18] by proposing a hierarchical Bayesian topic model. [7] made use of semi-supervised PLSA to integrate opinions between formatted text and blog text. [5] conducted a quantitative comparison between blog stream and news stream based on phrase graph. The nature of Twitter makes our work more difficult than these previous studies because tweets are short messages and different from traditional documents. In addition, no previous work has done comparison of topics in different views, i.e. topics of differen categories and topics of different types.

A most recent piece of work [13] 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.

7 Acknowledgement

This work was done during Xin Zhao's visit the group of Assistant Prof. Jing Jiang in Singapore Management University. We thank Jianshu Weng, Jing He and Prof. Ee-Peng Lim for insightful discussions. This is the long version of

our ECIR paper [19]. For NYT news dataset and Twitter-LDA code[4], please contact Wayne Xin Zhao via email `batmanfly@gmail.com` .

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5228–5235, 2004.

[3] L. Hong and B. D. Davison. Empirical study of topic modeling in Twitter. In *Proceedings of the SIGKDD Workshop on Social Media Analytics*, 2010.

[4] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, 2010.

[5] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[6] P. Li, J. Jiang, and Y. Wang. Generating templates of entity summaries with an entity-aspect model and pattern mining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 640–649, July 2010.

[7] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *Proceeding of the 17th international conference on World Wide Web*, pages 121–130, 2008.

[8] M. Paul and R. Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417, 2009.

[9] S. Petrović, M. Osborne, and V. Lavrenko. The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, 2010.

[10] D. Ramage, S. Dumais, and D. Liebling. Characterizing micorblogs with topic models. In *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*, pages 130–137, 2010.

[4] This piece of code is developed in a Java-based package and dependent on other classes in this package. We couldn't share the whole package but only the Twitter-LDA code. It can be easily modified to be reused.

[11] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, 2010.

[12] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, 2004.

[13] J. Teevan, D. Ramage, and M. Morris. #Twittersearch: A comparison of microblog search and web search. In *Proceedings of the fourth ACM WSDM*, 2011.

[14] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th International Conference on World Wide Web*, pages 111–120, 2008.

[15] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, pages 178–185, 2010.

[16] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proceedings of the Semantic Search 2010 Workshop*, 2010.

[17] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM International Conference on Web Search and Data Mining*, 2010.

[18] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 743–748, 2004.

[19] 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 *ECIR*, 2011.