

Uncovering Topic Dynamics of Social Media and News: The Case of Ferguson

Lingzi Hong¹, Weiwei Yang², Philip Resnik^{3,4}, and Vanessa Frias-Martinez^{1,4}

¹College of Information Studies, ²Department of Computer Science
³Department of Linguistics, ⁴Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA
{lzhong, wwyang, resnik, vfrias}@umd.edu

Abstract. Looking at the dynamics of news content and social media content can help us understand the increasingly complex dynamics of the relationship between the media and the public surrounding noteworthy news events. Although topic models such as latent Dirichlet allocation (LDA) are valuable tools, they are a poor fit for analyses in which some documents, like news articles, tend to incorporate multiple topics, while others, like tweets, tend to be focused on just one. In this paper, we propose Single Topic LDA (ST-LDA) which jointly models news-type documents as distributions of topics and tweets as having a single topic; the model improves topic discovery in news and tweets within a unified topic space by removing noisy topics that conventional LDA tends to assign to tweets. Using ST-LDA, we focus on the unrest in Ferguson, Missouri after the fatal shooting of Michael Brown on August 9, 2014, looking in particular at the topic dynamics of tweets in and out of St. Louis area, and at differences and relationships between topic coverage in news and tweets.

1 Introduction

The cascading activation model is a widely accepted model that explores the relationship among the government, the media, and the public [4]. The model explains how the framing of information extends down from the White House to the elites, the media, and then to the public. Information moves downward along the cascade with the framing of upper layers and becomes limited to highlights to the public. The structure emphasizes heavily the direction from the media to the public, given that historically the voice from the public has been comparatively weak.

However, social networks expand sources of information for every user and enable everyone to be a potential media source. Providing a public communication platform for everyone who is accessible to the Internet, social networks lead to increased participation in spreading information, expressing opinion, and public activism [7, 22]. During the Arab Spring, for example, Twitter promoted protest mobilization through reporting of real-time events and providing a basis for collaboration and emotional mobilization [1]. Effing et al. [3] show that political participation has been democratizing because of social media such as Facebook and Twitter, which enable more followers to engage in campaigns. So, in contrast to the traditional cascading activation model, the public may be gaining influence because of social networks.

This leads us to think about several questions. Can we observe the complex relations between the opinions of the public and the media? What does the public focus on, given highlighted topics spread by the media in an event? Are there any topics being followed by the public but not mentioned by the media? Specifically, we want to figure out what the media reports, the subjects of public attention, and the relation with and difference between these two. It is also important to observe that usually along with the evolution of an event, the topics of media and the public change over time. For example, after a gunshot accident, several relevant topics co-exist; meanwhile, the main topic may change from the description of the accident to the motivations, effect of the accident, and then to discussion about gun regulations. The changing topics form topic dynamics. By modeling the topic dynamics of media and the public, we can gain insight into the similarity, differences, and possible relationships between media topics and public topics.

In this study, we take the Ferguson unrest event of 2014 as an example, and analyze news and tweeted topics along with the unfolding of events. To make topics of news and tweets comparable, we propose a Single Topic LDA (ST-LDA) model to bring news and tweets under a unified framework, in which every tweet has only one topic, but news has a distribution of multiple topics, a novel approach that takes into account the length limitation of tweets and the greater complexity of news stories. The ST-LDA model tends to outperform LDA by removing noisy topics that conventional LDA tends to assign to tweets in a mixed collection of long and short documents. We explore our research questions by examining the shift of focus in news and tweets, specifically on the possible relation to burst, emergence, and decay of certain topics, the difference between topic dynamics of news and tweets, and whether there is strong influence of media on the public.

The contributions of this study have two main aspects:

1. We solve the technical problem of building topic models for a mixture of short and long documents. Conventional topic models such as LDA and PLSA [11] perform badly because co-occurrence patterns in short text are sparse. Our model considers the words in a tweet as a whole and assigns only one topic to a tweet, so that the main topic is more likely to be assigned. The evaluation results show that ST-LDA improves interpretability over LDA by 14%.
2. We bring the cascade activation model into a social media environment, reexamine the focus of media and the public in the Ferguson case, and bring new understanding of the influence of the media and formation of public opinions in social media.

2 Related Work

2.1 Detection of Topics in News and Social Media

The first and foremost step for comparing topics is the detection of topics in news and tweets. Tracking memes on the Internet [20] is one way to understand online information content. Memes are entities that represent units of information at the desired level of detail. The semantic units serve as clear clues for detecting dynamic change of diverse topics. However, in this approach only repeated topics can be detected.

Topic detection in tweets has been a challenge because of the short length [24]. Aggregating short tweets into a long document, such as author-based aggregation [23, 25], grouping by time slices [16], and by words [12] are ways to alleviate the problem. The bitern topic model [24] directly simulates the generation of word co-occurrence patterns in a corpus, and thus leads to more coherent topics. The Word Network Topic Model [26] also uses a word co-occurrence network to solve the sparsity problem. Cataldi et al. [2] detect emerging topics on Twitter by evaluating the life cycle of Twitter terms and user authority.

To train news and tweets together, Hu et al. [13] propose a joint Bayesian model for events transcripts and tweets. It assumes that event information can impose topical influence on tweets. Gao et al. [6] create a joint topic model to extract important and complementary pieces of information across news and tweets, and generate complementary information from both.

In these models, each tweet still has a distribution of topics, despite the fact that, given Twitter’s length limitation, a tweet is unlikely to have multiple topics. In contrast, we propose a model that assigns only one topic to each tweet and trains on tweets and news under a unified framework.

2.2 Topic Dynamics

Topic dynamics characterize the shift of topic proportions in a daily window. Dynamics of topics have been studied a great deal in research on the development of scientific areas [18], burst topics in publications [8], and public opinions on social media [10]. Morinaga and Yamanishi [19] employ a finite mixture model to recognize the emergency, growth, and decay of each topic in a system. Iwata et al. [15] build a sequential topic model to detect topic dynamics of document collections with multiple timescale. All these studies involve a single source of data, defining the calculation of topic dynamics in different ways. In our study, we bring tweets and news into a unified topic space so that we can compare the topic dynamics of news and tweets.

2.3 Comparison of Social Media and News

We aim to compare the topics between media and the public, a subject that has been studied both qualitatively and quantitatively. Sayre et al. [21] manually analyze thousands of videos and news media on Proposition 8 in California, and find that the post content in open social media reflects mainstream news, while posts also have influence on professional media coverage. Together they form opinion interactions between media and the public. However, the study required a large amount of human effort, and it is difficult to identify topics’ weight change during the evolving process. Hua et al. [14] explore the semantic and topical relationships between news and social media to reveal topic influences among multiple datasets. However, they focus on the influence between topics based on word probability, ignoring the time element of tweets. Leskovec et al. [17] introduce a meme-tracking technique to track topic shifts in news and blogs and observe a 2.5-hour lag between peaks of attention of a phrase in the news media

and in blogs, suggesting possible media influence on individuals; however their characterization of memes is limited to variants of quotations rather than a broader notion of topic.

Zhao et al. [25] propose Twitter-LDA, which assigns one topic to each tweet, however its premise is that each Twitter user has a distribution of topics and the topic of each tweet is drawn from its author’s topic distribution. Moreover, when they compare tweets and news media, they apply topic models separately and manually label topics for news and tweets. Topics in news and tweets with the same labels are compared. Although the meaning of topics is similar, the word distributions are actually different. This can make sense when considering all topics including arts, event, sports, etc., but the comparison is unlikely to be accurate enough for topics all focused on a single event. The Twitter-LDA approach cannot be compared to the proposed ST-LDA directly since it requires a large number of tweets per individual, which is typically not available when the tweet collection is done for events where millions of users might have just a few tweets each.

3 Methods

3.1 Dataset

We collected 13,238,863 tweets from August 10, 2014, to August 27, 2014 that contain the keyword “Ferguson” using the Twitter Streaming API. Since media may have a different influence on people who have experienced an event versus people who have not, and since perceptions of social events are usually affected by distance [5], we take geographic influence into consideration by using geo-tagged tweets as a sample of all tweets for content analysis; of the full set we collected, 110,280 (0.83%) are geo-tagged.¹ Previous work [9] shows that temporal patterns of tweet volume do not differ significantly between geo-tagged and non-tagged tweets, nor do the proportions of more and less influential users.

It is noteworthy that the media play their role in various ways, such as news reports, TV programs, and even accounts in different social networks. We identified news stories via the links published by 108 media accounts on Twitter, e.g. Washington Post, NBC News, ABC7News, etc, looking at all the tweets they published during the Ferguson event and identifying news reports from the links.² In total, we collected 1,338 news reports dated from August 11 to 27.

The same preprocessing is applied to the news and Twitter corpora, including tokenization, lemmatization, bigram detection, and removal of stop words, low frequency words, and high frequency words.³ After preprocessing and removing empty documents, the final corpus contains 1,275 news documents and 81,553 tweets.

¹ To identify locations of tweets, we refer to the geographic boundary file of 2014 TIGER/Line, <https://www.census.gov/geo/maps-data/data/tiger-line.html>

² Tweets from these media sources are filtered from our Twitter data.

³ News tokenization is done by OpenNLP, <https://opennlp.apache.org/>. Tweet tokenization is done by Twokenizer, <http://www.cs.cmu.edu/~ark/TweetNLP/>.

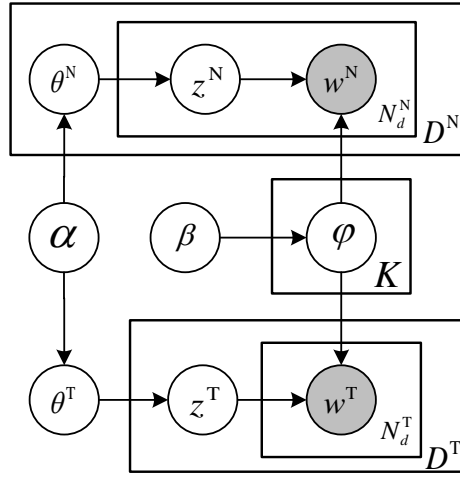


Fig. 1. Graphical Model of ST-LDA

3.2 Single Topic LDA

We introduce ST-LDA to jointly model short documents like tweets and long news documents.⁴ The key intuition is that a very short document like a tweet is unlikely to be related to multiple topics; therefore it can be modeled as having all its words generated from a single topic. In contrast, long documents like news stories are likely to follow conventional LDA assumptions, so each document is modeled conventionally as having a mixture of topics. At the same time, news and tweets are likely to discuss the same events in the world, so they share the same topic-word distributions. Figure 1 shows the graphical model of ST-LDA, where superscripts N and T denote news and tweets, respectively. The corresponding generative process of ST-LDA is as follows.

1. For each topic $k \in \{1, \dots, K\}$
 - (a) Draw word distribution $\phi_k \sim \text{Dirichlet}(\beta)$
2. For each (long) news document $d \in \{1, \dots, D^{\text{News}}\}$
 - (a) Draw a topic distribution $\theta_d^{\text{News}} \sim \text{Dirichlet}(\alpha)$
 - (b) For each token $t_{d,n}^{\text{News}}$ in news document d
 - i. Draw a topic $z_{d,n}^{\text{News}} \sim \text{Multinomial}(\theta_d^{\text{News}})$
 - ii. Draw a word $w_{d,n}^{\text{News}} \sim \text{Multinomial}(\phi_{z_{d,n}^{\text{News}}})$
3. Draw tweet background topic distribution $\theta^{\text{Tweet}} \sim \text{Dirichlet}(\alpha)$
4. For each (short) tweet document $d \in \{1, \dots, D^{\text{Tweet}}\}$
 - (a) Draw a topic $z_d^{\text{Tweet}} \sim \text{Multinomial}(\theta^{\text{Tweet}})$
 - (b) For each token $t_{d,n}^{\text{Tweet}}$ in document d
 - i. Draw a word $w_{d,n}^{\text{Tweet}} \sim \text{Multinomial}(\phi_{z_d^{\text{Tweet}}})$

⁴ Code is available at https://github.com/ywwbill/YWWTtools#st_lda_cmd

The key to the model is the combination of conventional LDA for news and the adjusted single-topic model component for the tweets. Different from LDA, the coverage of the word plate and document plate are different, adapted according to our assumptions. First, the word plate of ST-LDA’s tweet part (subscript N_d^T) only covers \mathbf{w} , which means that every word in a tweet is generated from the same topic. In LDA, the corresponding plate covers both \mathbf{z} and \mathbf{w} , denoting that every word has its own topic assignment and every document consists of a mixture of topics. Second, every tweet only has one topic; θ^T is outside the document plate (subscript D^T) and denotes a background topic distribution of tweets.

3.3 Posterior Inference

The Gibbs sampling equation for news documents is the same as conventional LDA. The probability of tweet d being assigned a topic k is computed as

$$\Pr(z_d = k | \mathbf{z}_{-d}, \mathbf{w}) \propto (N_k^{-d} + \alpha) \frac{\prod_{v=1}^V \prod_{i=0}^{N_{d,v}-1} (N_{k,v}^{-d} + \beta + i)}{\prod_{i=0}^{N_{d,\cdot}-1} (N_{k,\cdot}^{-d} + V\beta + i)}, \quad (1)$$

where N_k denotes the number of documents assigned to topic k ; $N_{d,v}$ is the count of word v in document d ; $N_{k,v}$ denotes the count of word v assigned to topic k . Marginal counts are denoted by \cdot^{-d} denotes that the count excludes document d .

3.4 Topic Dynamics

The output of ST-LDA can be used for further discovery of topic dynamics in tweets and news. Topic dynamics are characterized here as the temporal change in topics using a daily sliding window. Assuming that every news document has the same impact and contributes equally to the total media environment, the topic proportion of day t is the average of topic probabilities of all news documents on that day:

$$\bar{\theta}_{t,k}^{\text{News}} = \frac{\sum_{d=1}^{D_t^{\text{News}}} \theta_{d,k}}{D_t^{\text{News}}}, \quad (2)$$

where D_t^{News} denotes the number of news documents on day t ; $\theta_{d,k}$ is topic k ’s proportion in document d .

In contrast, in ST-LDA each tweet d has only one topic z_d . Under the same assumption that each tweet contributes equally to the voice of the public, the aggregation of daily tweet topic proportion is calculated as

$$\bar{\theta}_{t,k}^{\text{Tweets}} = \frac{\sum_{d=1}^{D_t^{\text{Tweets}}} \mathbb{1}(z_d = k)}{D_t^{\text{Tweets}}}, \quad (3)$$

where D_t^{Tweets} denotes the number of tweet documents on day t and $\mathbb{1}(\cdot)$ is an indicator function.

Given $\bar{\theta}_t^{\text{News}}$ and $\bar{\theta}_t^{\text{Tweets}}$, where t varies from August 11 to 27, we can identify topic dynamics by the changing of daily topic proportions.

4 Quantitative Evaluation of ST-LDA

In this section, we evaluate ST-LDA and LDA quantitatively by performing topic identification task on both news and tweets.⁵ We split our datasets into training (90%) and test (10%) sets, both for news and tweets, and evaluate the quality of topics given by ST-LDA and LDA respectively. We first align the topics given by ST-LDA and LDA using KL-divergence. The KL-divergence of topic k_1 given by LDA and topic k_2 given by ST-LDA is measured as

$$T_{k_1, k_2} = \text{KL}(\phi_{k_1}^{\text{LDA}} \parallel \phi_{k_2}^{\text{ST-LDA}}) = \sum_{v=1}^V \phi_{k_1, v}^{\text{LDA}} \log_2 \frac{\phi_{k_1, v}^{\text{LDA}}}{\phi_{k_2, v}^{\text{ST-LDA}}}. \quad (4)$$

Then we manually summarize each topic with a label and have two annotators annotated each tweet and news in the test set with one of the labels. To be strict, annotators are required to annotate “other” if none of the labels fit well, especially on news, because both LDA and ST-LDA give a distribution of topics on news documents. Due to the large number of tweets, we sample 5% of test tweets for annotation. The annotation agreement rates are 71.7% (91/127) and 79.1% (322/407) for news and tweets, respectively. The lower agreement rate for news is due to the different opinions about the main topic, since each news document usually covers multiple topics.

We only use the data points with agreed annotations and measure the two models’ accuracies. Since LDA gives a probability distribution, we consider its output as the topic with the highest probability.

After experimenting with different numbers of topics, we report the best results with 10 topics in Table 1. Although LDA has higher accuracy in news topic identification, the values are quite close. However, ST-LDA improves the accuracy by 14% in assigning the main topics to test tweets, which demonstrates its efficacy.

Model	News	Tweets
LDA	0.637	0.388
ST-LDA	0.615	0.525

Table 1. Topic Identification Accuracies

5 Qualitative Evaluation of ST-LDA

In this section, we evaluate the quality of some of the topics assigned to news and tweets and explore their temporal evolution. Table 2 shows five matched topics of LDA and ST-LDA on mixed documents. We also set a baseline by running LDA on news, and match the topics to ST-LDA topics by KL-divergence. Common topics in LDA on news and

⁵ Note that ST-LDA will not outperform LDA on perplexity, since the words in a tweet are generated from the same topic. However, the sacrifice of perplexity brings improvement in topic identification.

Model (Corpus)	Topic Label	Top Words
LDA (N+T) ¹	Obama Talk ²	happen, i'm, make, thing, talk, situation, what's, what's_happen, bad, you're
	Protest	tear_gas, protester, arrest, fire, medium, rt, protestor, street, crowd, pd
	Racist	black, white, loot, protect, community, racist, stop, race, citizen, riot
	Curfew	missouri, state, obama, national_guard, call, curfew, mo, press, governor, nixon
	Pray	peace, pray, justice, stand, love, tonight, hope, stay, family, safe
ST-LDA (N+T)	Obama Talk	obama, president, law_enforcement, house, holder, make, story, post, include, community
	Protest	tear_gas, arrest, protester, fire, rt, reporter, medium, shoot, crowd, street
	Racist	black, white, make, race, america, obama, stop, happen, situation, riot
	Curfew	missouri, curfew, state, national_guard, governor, nixon, call, gov, order, make
	Pray	peace, pray, stand, justice, night, love, tonight, today, family, safe
LDA (N)	Obama Talk	obama, president, house, make, white, news, national, deal, run, defense
	Protest	st_louis, nixon, protester, shooting, county, justice, aug., investigation, state, thursday
	Racist	black, make, white, cop, time, don't, year, good, man, thing
	Curfew	protester, johnson, tear_gas, crowd, curfew, night, fire, street, missouri, shoot
	Pray	(No matching topic)

¹ N: News. T: Tweets.

² This topic matches *Obama Talk* in ST-LDA according to KL-divergence. However, the topic label is “question of the situation”. For comparison we still name the topic *Obama Talk*.

Table 2. Topic Examples

ST-LDA on news and tweets show that ST-LDA keeps topics from news. Meanwhile the topic *Pray* only exists in the results of LDA and ST-LDA on mixed documents, which means *Pray* mainly exists in tweets. In addition, Twitter words such as *rt* and *gov* appear in the top words of topics given by ST-LDA. Therefore, ST-LDA is not biased to tweets or news and is able to discover topics from both.

Table 3 lists seven tweet examples and Table 4 shows their topic distributions. The topics are matched and numbered from 0 to 9.

The first three tweets' main topics given by LDA are the same as those given by ST-LDA. Although the main topics are consistent with the content, LDA assigns some probability to other topics. For example, though Tweet 1 contains words like *shoot*, it is not appropriate to assign this tweet to *Shooting Incident* which has top words like *street* and *Michael Brown*. Tweet 2 mainly talks about *Race and Community*, but LDA assigns the topic *Obama Talk* with probability 0.373 and small probabilities to other topics, which makes the main topic *Racism* only take 0.555.

Tweet 4 is an example in which LDA assigns the highest probability to multiple topics, namely, *Protest*, *Michael Brown*, *Shooting Incident*, *Emotion*, and *Race and Community*. In this situation, this tweet has no main topic. Tweet 5 is a case where ST-LDA assigns the right topic but LDA fails.

Tweets 6 and 7 fit in none of the ten topics, i.e. their labels are “other”. Tweet 6 seems to have no clear topic. Tweet 7 talks about medical care for injuries, which is not a main topic discovered by either ST-LDA or LDA.

No.	Content
1	“@bkesling: “Hands up, don’t shoot” after tear gas fired in #Ferguson http://t.co/9zQlh31wQg ” modern day America... #PrayForFerguson
2	80% black folks think #Ferguson raises “important issues about race that need to be discussed,” only 37% of white folks do. Very sad.
3	You guys can’t blame that cop in #Ferguson. Shooting your gun 6 times is literally the answer to every question in their training manual.
4	#fergusongate media get it straight. U act like those who don’t live in ferguson can’t protest. This is for all blacks everywhere.
5	But thank God for social media though. Imagine if we’re dependent on the news to tell the “truth” about what’s really happening in #Ferguson
6	@MikeHolmzy that’s true. But I’m just talkin about ferguson
7	County will not pay medical bills for toddler hurt in... http://t.co/8k8Hee5B63 via @sharethis #ferguson can you believe GA is doing this?

Table 3. Tweet Examples

Tweet Numbers		1	2	3	4	5	6	7	
LDA Topic Distribution	0	Obama Talk	0.017	0.373	0.011	0.017	0.888	0.025	0.017
	1	Protest	0.517	0.009	0.011	0.183	0.013	0.025	0.017
	2	Racism	0.017	0.555	0.233	0.017	0.013	0.025	0.350
	3	Curfew	0.017	0.009	0.011	0.017	0.013	0.025	0.350
	4	Michael Brown	0.017	0.009	0.567	0.183	0.013	0.025	0.017
	5	News Report	0.017	0.009	0.011	0.017	0.013	0.025	0.017
	6	Pray	0.017	0.009	0.011	0.017	0.013	0.025	0.017
	7	Shoot Incident	0.350	0.009	0.011	0.017	0.013	0.025	0.017
	8	Emotion	0.017	0.009	0.122	0.183	0.013	0.775	0.017
	9	Race and Community	0.017	0.009	0.011	0.183	0.013	0.025	0.183
ST-LDA Topic		1	2	4	8	5	1	6	

Table 4. Tweet Topic Comparison. Content of the tweets can be referred in Table 3. The topics given by LDA and ST-LDA are matched.

Next, we perform a qualitative evaluation of the topic dynamics in news and tweets provided by LDA and ST-LDA. Figures 2(a) and 2(b) show the changes in news topic proportions from August 11 to 27 according to LDA and ST-LDA, respectively.

The topic distribution given by LDA (2(a)) is highly skewed toward two main topics—*Shooting Incident* and *Race and Community*. Other topics take small proportions, so it is hard to identify their proportion changes. Meanwhile ST-LDA yields results that are slightly better in representing different topics. *Obama Talk* is discovered as a main topic. It keeps a relatively stable proportion of 20%, and peaks after some important events related to Obama. On August 12, Obama addressed the shooting and urged the Ferguson community to stay calm. On August 14, he gave a talk saying that there is no excuse for protests to turn into violence, which leads to the peak of topic *Obama Talk* on August 14 and 15. This demonstrates that the topics detected by ST-LDA are consistent with important events in the timeline.

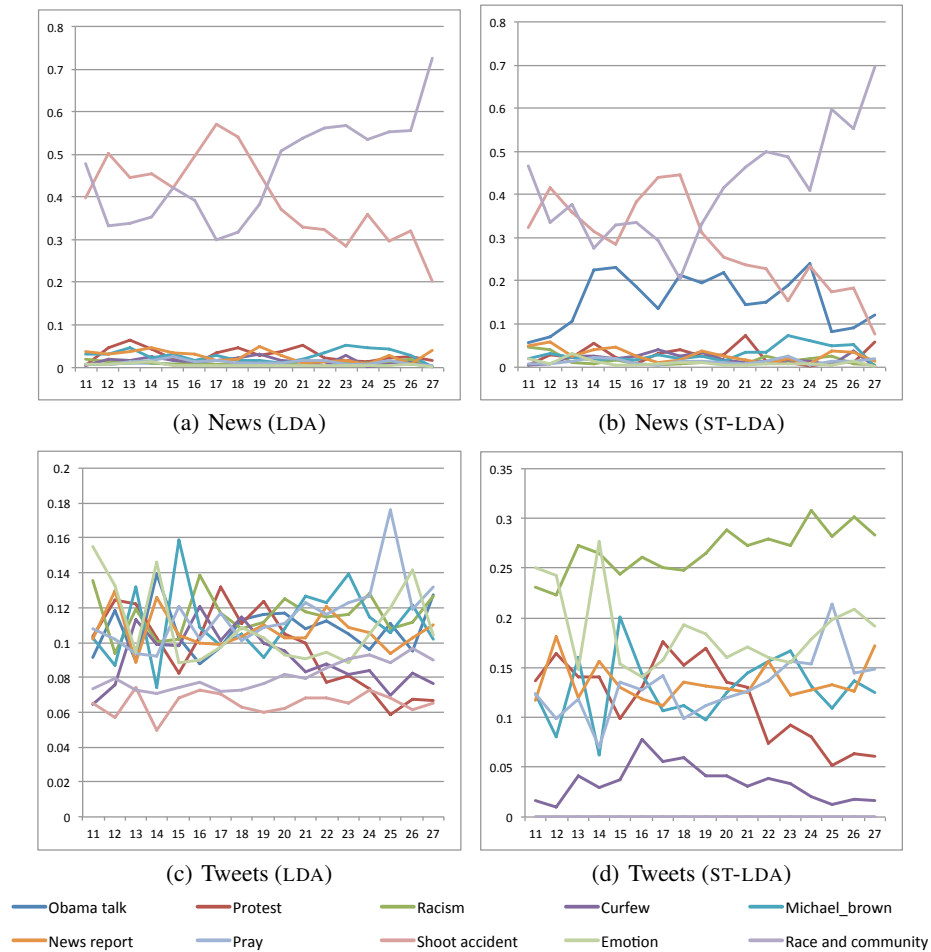


Fig. 2. Topic Dynamics of News and Tweets by LDA and ST-LDA

There is more variance in topic dynamics of tweets according to ST-LDA than LDA, as shown in Figures 2(c) and 2(d). The proportions of topics are close to each other in topic dynamics according to LDA, which makes it hard to identify the main topics for each day. In comparison, ST-LDA gives results with more variation of topics along the timeline. It is clear that after the shooting incident, *Emotion* of the public surges to a peak on August 11. After the protest event, another *Emotion* topic appears on August 14. Meanwhile, the proportion of *Pray* topic stays relatively stable from August 11 to August 24, but increases a great deal on the day when Michael Brown's funeral is held.

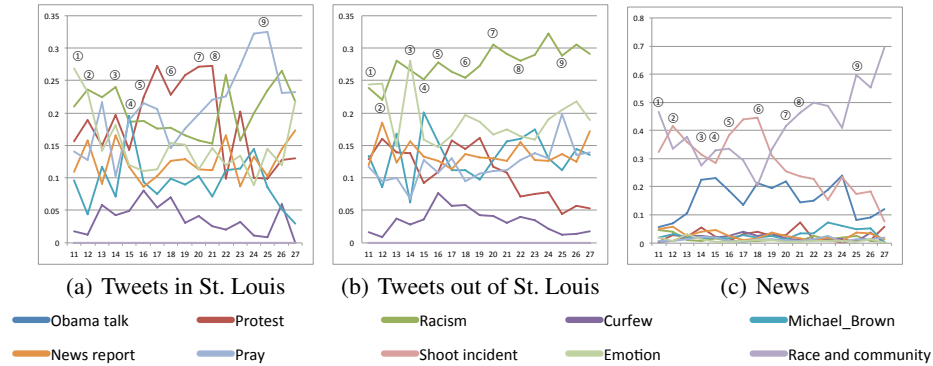


Fig. 3. Topic Dynamics of Tweets and News by ST-LDA. Important events: ① Aug 11: Unrest continued; ② Aug 12: First Obama talk; ③ Aug 14: Second Obama talk and Nixon announced law enforcement operation; ④ Aug 15: Robbery video was released; ⑤ Aug 16: Curfew was imposed; ⑥ Aug 18: National Guard was deployed; ⑦ Aug 20: A grand jury convened to begin determining of crime and streets become quiet; ⑧ Aug 21: National Guard withdrew; ⑨ Aug 25: Michael Brown’s funeral.

6 Topic Dynamics of Tweets and News

In this section, we use ST-LDA to analyze the Ferguson event. Considering the effect of distance on event perception [9], we analyze the topic dynamics for tweets in and out of St. Louis, where the Ferguson unrest took place. First, we compare the general topic dynamics to ground truth events to see the different focuses of the media and the public. Then, we compare the topic dynamics of news and tweets in and out of the St. Louis area to explore possible relations between the media and the public. For evaluation and comparison, we use the timeline of important events since Michael Brown’s death. It includes information from different perspectives: shooting incident, looting, FBI investigation, Obama talk, protests, curfew, Michael Brown’s funeral and so on.

6.1 Tweet Topics in and out of the St. Louis Area

The tweet topic dynamics in and out of the St. Louis are shown in Figure 3. Topic dynamics of tweets are highly related to ground truth events.

Topics *Curfew*, *News Report*, and *Michael Brown* share similar change patterns for tweets both in and out of the St. Louis area. *Curfew* increases to a peak on August 16 when Governor Nixon declared a state of emergency and imposed a curfew (see ⑤). It peaks at around 5%, indicating that it is not the main issue of the public. The topic *Pray* shares similar dynamics with a large increase of tweets on this topic on August 25 when Michael Brown’s funeral is held (see ⑨). However, more than 35% of tweets in St. Louis are about *Pray*, while outside of St. Louis it is 20%.

In the topic *News Report*, top words, such as *news*, *watch*, *live*, *report*, *coverage*, and *rt* indicate that the topic is mainly about the description and citation of information from news and TV. This is direct evidence of media influence on tweets. Tweets in and

out of St. Louis have similar topic dynamics for *News Report*, indicating that people in and out of St. Louis paid similar attention to *News Report*.

However, differences in the topic dynamics of tweets show different perceptions of events for people in and out of the St. Louis area. More tweets in St. Louis talk about *Protest*, while out of St. Louis more tweets talk about *Racism*. From August 18, when Governor Nixon deployed the National Guard to Ferguson (see ⑥), to August 21, when the National Guard withdrew (see ⑧), protests and conflicts kept occurring. People in the Ferguson area are closer and more connected to the protests, so tweets with this topic surge to take more than 25% of the volume. Meanwhile, the proportion of *Protest* tweets outside St. Louis is far lower. We surmise that members of the public who are not involved in the event tend to have less focus on the precise situation on the ground, and therefore are more likely to discuss the situation abstractly; thus *Racism* takes the majority most of the time.

The dynamics of the *Emotion* topic also differ geographically. Michael Brown was killed on August 9, and anger is the major topic of tweets in St. Louis; then *Emotion* tweets keep decreasing, and only take up 10%–15% of the volume. However, outside the St. Louis area, there is a lag effect of the *Emotion* explosion on August 14. One possible reason for this is that news takes time to spread and the public outside Ferguson needs more information to understand what happened. Another possibility is that segments of the news media emphasize emotion because in the media business stories connected with strong negative emotions attract attention, which is good for business; consider the old adage “if it bleeds, it leads”.

It is also worth noting that the proportion of each topic outside St. Louis changes less compared to tweets in St. Louis. Although tweets for certain topics may increase in some time, the proportion of tweets in each topic keeps relatively steady. It is possible that people outside St. Louis have different sources of information like news and social media, so their focus is more dispersed.

6.2 News Topics

As shown in Figure 3, there are three main topic lines in news, *Obama Talk*, *Shooting Incident*, and *Race and Community*, which do not exist in tweets. According to the top words in *Shooting Incident* (*shoot, protester, michael_brown, st_louis, tear_gas*, etc.), it is quite similar to the topic *Michael Brown* (*shoot, kill, officer, unarmed, michael_brown*, etc.), which takes a certain proportion in tweets, and some proportion in news. Although news and tweets are talking about the same topic, the words they use are quite different, which leads to different topic assignments by ST-LDA. Similarly, the *Racism* topic exists mostly in tweets, while *Race and Community* mainly exists in news. These two topics are both about racism and human rights, but there is little overlap of tweets and news in the two topics. One possible reason is that the language in tweets is closer to spoken language, while news uses more formal written language; another is that media and the public frame the same event differently. According to the top words in two topics, there are more negative words in *Racism* such as *stop* and *riot*. In the topic *Race and Community*, words like *make, good*, and *community* are indicators of positive emotion. It seems possible that while the public tended to display negative emotion about race

issues during the Ferguson unrest, the media tried to describe and lead the discussion in a constructive way.

Among the main topics, only *Obama Talk* is related to ground truth events. The proportion of topic *Obama Talk* increases from August 12, when Obama first addressed the shooting (see ②), reaching a peak on August 14 when Obama addressed the situation in Ferguson again (see ③). After that, the proportion stays steady at about 20%. Two weeks after the shooting incident, this topic then decreases.

Of the minor topics, only *Curfew* is closely related to the occurrence of certain events. The emergence of *Curfew* in the news appears right after the day when Governor Nixon declared the curfew (see ⑤). The topic *Protest* has two peak points on August 14 and 21, which are the start and end dates of the National Guard deployment, respectively (see ③ and ⑧).

Topics *Pray* and *Emotion* take a very small part (less than 5%). It seems reasonable that tweets are more subjective and contain more words about feelings, emotions, and praying, while news is more serious and objective, avoiding emotional leading. Despite the explosion of *Emotion* and *Pray* in tweets, there is no corresponding burst of the topic in news. This may indicate that such emotional changes in the public are not reflected in news, or that news reacts to the emotions with other topics. The relation is hard to determine with certainty and it is not clear which news topics might be in reaction to public emotions.

6.3 Influence of News on Tweets

From the comparison of topic dynamics in tweets and news, we find that tweets have more diverse topics, while news appears to have only three main themes. Topics related to Obama maintain a stable proportion in news reports, and the investigation report and discussion of race issues keep alternating dominance. On the other hand, tweet topics are more diverse and change along with the evolution of the event.

In news and tweets, both *Racism* and *Shooting Incident* are discussed quite a bit, but in different ways. In tweets, the topic *Michael Brown* takes the majority, while in news *Shooting Incident* is more dominant. Top words in *Michael Brown* mainly include *shoot*, *kill*, *officer*, *unarmed*, and *michael.brown*, which reflects the public paying more attention to describing the triggering incident. However, top words in *Shooting Incident* are *shoot*, *protester*, *michael.brown*, *st.louis*, and *tear_gas*, which all seem to reflect more of a big picture of the larger series of incidents. Meanwhile, as we have noted, discussions of race appears to be framed more positively or constructively in news than in tweets.

The above results seem to support the theory of the cascade model [4] from the perspective of the role media plays. In addition, the existence of the *News Report* topic shows that the public accepts information from the media. Meanwhile, the topic dynamics of tweets in and out of the St. Louis area show the possible influence. There is a lag effect of the *Emotion* explosion, and there is much more discussion of *Racism* among people in St. Louis, which corresponds to the major topic of *Race and Community* in news.

However, unlike the idea of a cascade leading the public to focus on what the media focuses on, topics for the two are still quite different. Under the influence of media,

tweets do not simply repeat topics of news. For instance, the discussion of *Race and Community* and *Obama Talk* is rare in tweets. In addition, people on social media show more emotional change, which can be reflected by the dynamics of the *Emotion* and *Pray* topics when certain events happened. One might argue that this provides evidence of publics displaying primarily emotional rather than rational responses.

In summary, the topic dynamics of news and tweets in the Ferguson case form a picture in which:

1. the media tends to have a smaller set of topics that they emphasize consistently in coverage, in contrast to public opinions, which are more diverse and subject to change with new events;
2. both news and tweets describe and discuss the event; however, the news tends to link events together, while tweets tend to have a quicker response to events;
3. in the context of social media, the public tends to generate information not necessarily following the news; specifically, people in St. Louis prefer to report their experiences on Twitter by quickly responding to events such as protest, while some hot topics in the news did not seem to attract much attention on Twitter;
4. emotion tweets, such as *Emotion* and *Pray*, take a significant part; however, media reflection of these topics seems to be rare.

7 Discussion and Future Work

We introduced a new topic model, ST-LDA, that brings news and tweets under a unified topic space, so that topics of news and tweets are comparable. At the same time, it provides a solution to finding common topics from a mixed collection of long and short documents. The results show that ST-LDA is able to detect common topics in tweets and news and assign the main topic to each tweet more accurately. We plan to extend ST-LDA so that it can handle a wider range of document types.

Our analysis of dynamics showed how news and Twitter users reacted to the Ferguson case. However, it still remains to be seen whether our results generalize to other situations. Are there cases where the media and the public have different patterns of reaction? Moreover, we only used tweets and news over a limited time window and only looking at short-term influence. What about the long-term effects? Is it possible to track opinions on events like Ferguson even long after the events? We hope to address these questions in future work.

Acknowledgement

We thank anonymous reviewers for their insightful comments.

References

1. Breuer, A., Landman, T., Farquhar, D.: Social media and protest mobilization: Evidence from the Tunisian revolution. *Democratization* (2014)

2. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of Conference on Knowledge Discovery and Data Mining (2010)
3. Effing, R., van Hillegersberg, J., Huibers, T.: Social media and political participation: are Facebook, Twitter and YouTube democratizing our political systems? In: International Conference on Electronic Participation (2011)
4. Entman, R.M.: Framing: Towards clarification of a fractured paradigm. *Journal of Communication* (1993)
5. Fujita, K., Henderson, M.D., Eng, J., Trope, Y., Liberman, N.: Spatial distance and mental construal of social events. *Psychological Science* (2006)
6. Gao, W., Li, P., Darwish, K.: Joint topic modeling for event summarization across news and social media streams. In: Proceedings of the ACM International Conference on Information and Knowledge Management (2012)
7. González-Bailón, S., Borge-Holthoefer, J., Rivero, A., Moreno, Y.: The dynamics of protest recruitment through an online network. *Scientific Reports* (2011)
8. He, D., Parker, D.S.: Topic dynamics: An alternative model of bursts in streams of topics. In: Proceedings of Conference on Knowledge Discovery and Data Mining (2010)
9. He, J., Hong, L., Frias-Martinez, V., Torrens, P.: Uncovering social media reaction pattern to protest events: A spatiotemporal dynamics perspective of ferguson unrest. In: International Conference on Social Informatics (2015)
10. He, Y., Lin, C., Gao, W., Wong, K.F.: Tracking sentiment and topic dynamics from social media. In: Proceedings of International Conference on Weblogs and Social Media (2012)
11. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (1999)
12. Hong, L., Davison, B.D.: Empirical study of topic modeling in Twitter. In: Proceedings of Conference on Knowledge Discovery and Data Mining (2010)
13. Hu, Y., John, A., Wang, F., Kambhampati, S.: ET-LDA: Joint topic modeling for aligning events and their Twitter feedback. In: Proceedings of the Association for the Advancement of Artificial Intelligence (2012)
14. Hua, T., Yue, N., Chen, F., Lu, C.T., Ramakrishnan, N.: Topical analysis of interactions between news and social media. In: Proceedings of the Association for the Advancement of Artificial Intelligence (2016)
15. Iwata, T., Yamada, T., Sakurai, Y., Ueda, N.: Sequential modeling of topic dynamics with multiple timescales. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2012)
16. Lau, J.H., Collier, N., Baldwin, T.: On-line trend analysis with topic models: #Twitter trends detection topic model online. In: Proceedings of International Conference on Computational Linguistics (2012)
17. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of Conference on Knowledge Discovery and Data Mining (2009)
18. Mane, K.K., Börner, K.: Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences* (2004)
19. Morinaga, S., Yamanishi, K.: Tracking dynamics of topic trends using a finite mixture model. In: Proceedings of Conference on Knowledge Discovery and Data Mining (2004)
20. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., Menczer, F.: Detecting and tracking the spread of astroturf memes in microblog streams. *arXiv preprint arXiv:1011.3768* (2010)
21. Sayre, B., Bode, L., Shah, D., Wilcox, D., Shah, C.: Agenda setting in a digital age: Tracking attention to California Proposition 8 in social media, online news and conventional news. *Policy & Internet* (2010)
22. Tufekci, Z., Wilson, C.: Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication* (2012)

23. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: Finding topic-sensitive influential Twitterers. In: Proceedings of ACM International Conference on Web Search and Data Mining (2010)
24. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of World Wide Web Conference (2013)
25. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: Proceedings of the European Conference on Information Retrieval (2011)
26. Zuo, Y., Zhao, J., Xu, K.: Word network topic model: A simple but general solution for short and imbalanced texts. Knowledge and Information Systems (2014)