Social Media vs. News Media: Analyzing Real-World Events from Different Perspectives

Liqiang Wang[1], Ziyu Guo[1], Yafang Wang[1(✉)], Zeyuan Cui[1], Shijun Liu[1], and Gerard de Melo[2]

[1] Shandong University, Jinan, China
yafang.wang@sdu.edu.cn
[2] Rutgers University, New Brunswick, USA

Abstract. For a long time, the news media has played a crucial role not only as an information provider, but also as an influential source of opinion and commentary. Nowadays, platforms such as Twitter provide an alternative to the traditional one-way interaction, enabling users to voice their opinions. Hence, one can obtain a more comprehensive picture of the range of perspectives on real-world events by considering both news and social media sources. In this paper, we compare mainstream news and Twitter data on 18 well-known real-world events from six different categories. We propose the event-based authoring model (EvA), a novel probabilistic model to capture the content characteristics of an event with respect to aspect, category and background word distributions. These results allow us to analyze the real-world events in different perspectives.

Keywords: News media · Social media · Real-world event analysis
Topic model

1 Introduction

As online social media and online news continue to mature, increasing numbers of people rely on online media platforms to obtain information about the world as well as to express their personal opinions about various kinds of events. Such platforms are now among the primary sources that people rely on to keep track of current events in the world. Hence, online media possess unprecedented power to influence people's opinions. Clearly, there are substantial differences between social media and news media with regard to linguistic properties, distributions of opinions, sentiment, subjectivity, authenticity, immediacy, to name but a few. One study [4] determined that the value of news has remained constant, but that most raw content now is available both to journalists and to social media users. However, with the increased prominence of social media platforms, these

L. Wang and Z. Guo—Contributed equally.

S. Hartmann et al. (Eds.): DEXA 2018, LNCS 11030, pp. 471–479, 2018.
https://doi.org/10.1007/978-3-319-98812-2_43

now are also beginning to serve as gatekeepers on the news media. It has now become crucial to consider the interplay between social and news media, given the role that the two play in shaping the public's opinions on world events. Despite the abundance of research studying various aspects of online social media and of online news, this connection between the two has not received sufficient attention.

As more and more people participate in online discussions, analysts pay increasingly focus on mining the public perception, opinions, and online interactions pertaining to various real-world events, including, for instance, political events [8] and natural hazards [14]. One study [7] analyzed the engagement of Twitter users in response to real-world events. Although there have been numerous studies related to real-world event analytics, most of them focus on analyzing the behavior or online users rather than taking an event-centric perspective. To address this gap, the present study proposes a novel analytical model, considering event aspects as well as categories.

Both news and social media serve as vehicles for information authoring, dissemination, and diffusion. While there have been studies that sought to characterize specific aspects of how social and news media differ [11,15], in this paper, we attempt to provide a more comprehensive data-driven analysis of how news and social media differ, focusing on an event-centric perspective by highlighting how the two differ in covering the same set of events.

Our main contributions include:

1. We propose a new means of analyzing real-world events by accounting for both news and social media.
2. We introduce a new model called EvA to probabilistically model the events based on intuitions of how a journalist or user may compose an article.
3. We perform an analysis of event aspects to provide a contrastive analysis discriminating between news and social media.

2 Related Work

Event Detection. In this paper, we consider the somewhat rigid method of detecting real-world events based on keywords to compile our dataset. However, there are also several works that focus on extracting events using more involved methods. One study aims at related event discovery by extracting local events from web pages [10]. Certain previous work has also specifically aimed at detecting events in tweets [1]. For further references, the reader is referred to a recent survey that summarizes techniques for event detection in Twitter [2].

Event Analysis. In terms of data analysis, there has been previous work focusing on news and social media. ET-LDA [9] is a joint topic model for aligning events and corresponding feedback on Twitter. For news media, a recent method ranks the daily news events according to their importance [13]. Castillo et al. predict the life cycle of online news stories by analyzing reactions to the news on

social media [3]. For further information on the challenges and possible solutions for event analytics on social media, the reader is referred to a recent overview [6].

3 The EvA Model for Multi-perspective Event Analytics

3.1 Preliminaries and Definitions

To better understand the analytics and the model, we highlight some definitions in our model. **Events** refer to happenings in the world, e.g., the *Oscars*, *Nobel Prize*, or *terrorist attacks in France*. **An aspect** refers to a particular issue or subject that can be discussed about a given event. For instance, for our Academy Award event data, *Best picture*, *La La Land*, and *Emma Stone* are among the most trending aspects being discussed. **An event category** is a classification of events with regard to the event context, type, or attributes. For instance, based on the event context, we can consider categories such as *disasters*, *business*, *political events*, etc. **A background word** is an auxiliary word used by the authors to help in phrasing their thoughts, but typically does not express a specific opinion in the context under consideration, e.g., *actor*, *immigrate*, and *black*. In addition, the main notational conventions are enumerated in Table 1.

Table 1. The key notations used in this paper.

Notations	Description
E, D, T, C, N, B	Number of events ($e = 1 : E$), documents ($d = 1 : D$), aspects ($k = 1 : K$), event categories ($c = 1 : C$), words ($n = 1 : N$), and background word distribution
w, z, c	Word, aspect, category
x^0, x^1	Word indicators, one per word
ψ^0, ψ^1	Word bias on event category, aspect and background
$\lambda_0^0, \lambda_1^0, \lambda_0^1, \lambda_1^1$	Beta prior for hidden variables
$\theta, \phi, \sigma, \Omega$	Distribution of document-aspect, aspect-word, category-word and background-word
$\alpha, \beta, \delta, \pi$	Dirichlet prior for hidden variables

3.2 EvA Model Description

To address the aforementioned aims, we devise the EvA (Event-based Authoring) model in Fig. 1, inspired by the process of how a person – e.g., a journalist or blogger, or, alternatively, a social media user – authors a document, i.e., an article or a posting. Typically, the authoring process would be triggered by an event. We assume that the first decision that authors make is to settle on what event aspects they wish to write about. Because the purpose of writing a document is to express an opinion or to report on some aspect. Next, we assume that the

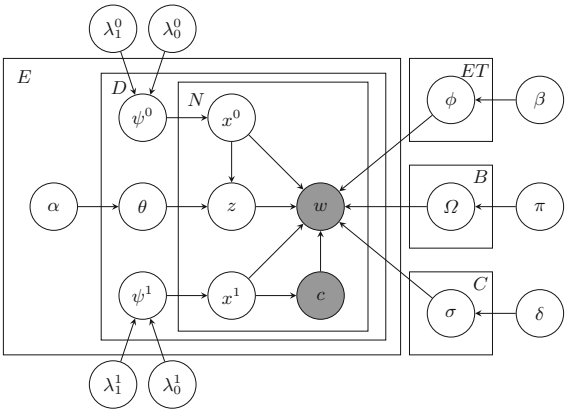


Fig. 1. EvA probabilistic graphical model

event category will exert special influence. For different category events, authors tend to follow different conventions while writing an article. Finally, authors will need to turn relevant information and thoughts on the various aspects of the event into a coherent narrative by organizing the words to form a suitable series of written sentences. For this, we distinguish three kinds of words: aspect words, category words, and background words. Background words, as mentioned, refer to auxiliary words, which can be thought of as being added last, as they merely assist in casting the thoughts and opinions into proper phrases.

Algorithm 1. The generation process for documents.

Draw background word distribution $\Omega \sim Dir(\pi)$
for each category $c = 1, ..., C$ **do**
  Draw category word multinomial distribution $\sigma_c \sim Dir(\delta)$
for each event $e = 1, ..., E$ **do**
  for each aspect $k = 1, ..., K$ **do**
    Draw aspect word multinomial distribution $\phi_{ek} \sim Dir(\beta)$
  for each document $d = 1, ..., D$ **do**
    Draw a Bernoulli distribution $\psi_d^0 \sim Beta(\lambda_1^0, \lambda_0^0)$
    Draw a Bernoulli distribution $\psi_d^1 \sim Beta(\lambda_1^1, \lambda_0^1)$
    Draw document aspect mixture $\theta_d \mid \alpha_e \sim Dir(\alpha_e)$
    for each word $n = 1, ..., N$ **do**
      Sample $x_n^0 \sim Bernoulli(\psi_d^0)$
      if $x_n^0 = 1$ **then**
        Sample an aspect $z_n \mid \theta_d \sim Multi(\theta_d)$
        Draw word $w_n \mid z_n \sim Multi(\phi_{ez_n})$
      if $x_n^0 = 0$ **then**
        Sample $x_n^1 \sim Bernoulli(\psi_{d_n}^1)$
        if $x_n^1 = 1$ **then**
          Draw word $w_n \mid c \sim Multi(\sigma_c)$
        if $x_n^1 = 0$ **then**
          Draw word $w_n \sim Multi(\Omega)$

3.3 The Inference Process

The posterior probability of the EvA model is as follows:

$$P(\theta_{1:D}, \phi_{1:K}, \sigma_{1:C}, \Omega, \psi^{\mathbf{0}}{}_{1:C}, \psi^{\mathbf{1}}{}_{1:K}, \mathbf{z}, \mathbf{x^0}, \mathbf{x^1} \mid \alpha, \beta, \delta, \pi, \lambda_0^0, \lambda_1^0, \lambda_0^1, \lambda_1^1, \mathbf{w}, \mathbf{c}) \quad (1)$$

Unfortunately, computing this posterior probability is intractable with all the variables. However, we can approximate it via collapsed Gibbs sampling. This involves integrating out the following hidden variables: $\theta_d, \phi_k, \sigma_c, \Omega, \psi_c^0$, and ψ_k^1.

We define C_{eq}^{EQ} as the number of times instance e appeared with instance q, e.g., C_{wk}^{WK} gives the number of times word w was assigned to aspect k. In addition, we use subscript $-i$ to denote the counting variable that excludes the i^{th} word index in the corpus. Moreover, $C_{\cdot w}^{BW}$ refers to the number of times that word w is sampled from the background word distribution. As a result, we finally obtain the following conditional posterior distribution:

$$P(x_n^0 = 1, z_n = k \mid w_n = w) \propto (C_{d1,-n}^{DX^0} + \lambda_1^0) \times \frac{C_{kw,-n}^{KW} + \beta}{\sum_{w'}(C_{kw',-n}^{KW} + \beta)} \times \frac{C_{dk,-n}^{DK} + \alpha_{ek}}{\sum_{k'}(C_{dk',-n}^{DK} + \alpha_{ek})} \quad (2)$$

$$P(x_n^0 = 0, x_n^1 = 1 \mid w_n = w, c_d = c) \propto (C_{d0,-n}^{DX^0} + \lambda_0^0) \times \frac{C_{d1,-n}^{DX^1} + \lambda_1^1}{C_{d1,-n}^{DX^1} + C_{d0,-n}^{DX^1} + \lambda_0^1 + \lambda_1^1} \times \frac{C_{cw,-n}^{CW} + \delta}{\sum_{w'}(C_{cw',-n}^{CW} + \delta)} \quad (3)$$

$$P(x_n^0 = 0, x_n^1 = 0 \mid w_n = w) \propto (C_{d0,-n}^{DX^0} + \lambda_0^0) \times \frac{C_{d0,-n}^{DX^1} + \lambda_0^1}{C_{d1,-n}^{DX^1} + C_{d0,-n}^{DX^1} + \lambda_0^1 + \lambda_1^1} \times \frac{C_{\cdot w,-n}^{BW} + \pi}{\sum_{w'}(C_{\cdot w',-n}^{BW} + \pi)} \quad (4)$$

α_e is a non-uniform vector related to the event e. Considering its simplicity and speed, we update α_{ek} according to $\alpha_{ek} = \frac{1}{N_e}\sum_d \frac{C_{dk}^{DK}}{C_{d\cdot}^{DK}}$ in each iteration of Gibbs sampling [12]. Assume that N_e is the number of documents in event e and document d belongs to event e.

4 Experiments and Analytics

4.1 Data Description

Our dataset includes 6 categories, and for each category we have 3 events. For each of these events, we compare the data from news and social media. The news media documents are sourced from the STICS project [5]. We rely on event keywords to filter these news articles so as to obtain related articles for a given event. The social media data is crawled from Twitter by relying on the Twitter API, using the same selection criteria. The overall statistics of the resulting dataset are given in Table 2.

Table 2. Dataset description

Category	ID	Keyword	Data period	Twitter	News
Armed conflicts and attacks	1	France attack	2016-06-14–2016-08-14	173,025	3,318
	2	Orlando shooting	2016-05-12–2016-07-12	525,891	3,059
	3	Turkish coup	2016-06-15–2016-08-15	145,952	2,111
Disasters and accidents	4	California wildfire	2016-07-16–2016-09-16	44,954	407
	5	Hurricane matthew	2016-09-03–2016-11-03	570,124	1,604
	6	Louisiana flood	2016-07-17–2016-09-17	148,952	353
Business and economy	7	Federal reserve	2016-11-15–2017-01-15	54,213	2,141
	8	OPEC oil	2016-11-10–2017-01-10	116,429	1,450
	9	Trump TPP	2016-12-23–2017-02-23	60,798	470
Politics and elections	10	Trump protest	2016-12-20–2017-02-20	307,263	3,009
	11	Trump inauguration	2016-12-20–2017-02-20	525,149	6,082
	12	Trump tax	2017-03-26–2017-05-26	395,460	3,093
Arts and culture	13	Grammys	2017-01-12–2017-03-12	532,296	465
	14	Oscars	2017-01-26–2017-03-26	326,714	1,088
	15	Nobel prize	2016-09-08–2016-11-08	297,890	1,407
Sports	16	Super bowl	2017-01-06–2017-03-06	288,166	2,511
	17	Olympics	2016-07-13–2016-09-13	511,042	5,816
	18	NBA finals	2017-05-06–2017-07-06	254,798	936

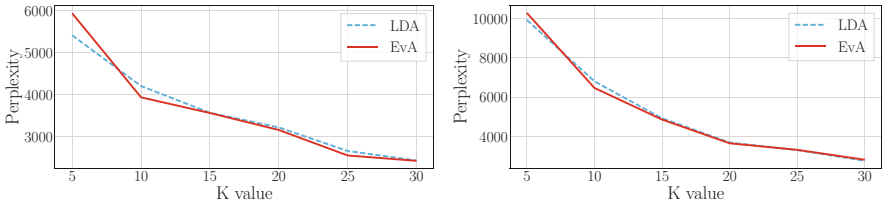
For data pre-processing, we remove stopwords and count the document (tweet) frequency for each word. We then remove words that appear in less than 10 documents for news and less than 20 tweets for Twitter. Finally, we obtain a vocabulary with 37,812 words for news media and 53,330 for Twitter.

4.2 Experiment

Experiment Configure. Regarding the parameters of our model, the prior for the hyper-parameters are set differently for news and Twitter. The values for δ, β, π are all set to 0.01 for news and 0.1 for Twitter, without further tuning, and λ_0^0, λ_1^0, λ_0^1, λ_1^1 are all set to 1, except that λ_1^0 set as 100 for news. The number of iterations for Gibbs sampling are fixed at 1000 for all methods. For baseline LDA, we set both α and β as 0.01 for news and 0.1 for Twitter.

Perplexity Comparison. We can rely on the perplexity as a measure to determine the most proper aspect number K. We thus present the results regarding the average word perplexity for different numbers of aspects. Given the estimated distribution q and the document set D_{test} for testing, we can compute the perplexity according to Eq. (5). A_d is the set of all aspect words in document d.

$$\text{perp}(D_{test} \mid q) = exp\{-\frac{\sum_{d \in D_{\text{test}}} \sum_{w \in A_d} logq(w \mid d)}{\sum_{d \in D_{\text{test}}} \mid A_d \mid}\} \tag{5}$$
$$q(w \mid d) = \sum_{k \in K} q(w \mid k) q(k \mid d)$$



(a) News (b) Twitter

Fig. 2. The perplexity results on news and Twitter with different K settings

The perplexity results are plotted in Fig. 2. The training and test splits are obtained by randomly selecting 80% documents as the training set, and reserving the rest for testing. From the results, we can observe that our model EvA outperforms the standard LDA. Considering that EvA also captures the category and background feature by separating words out from the aspect distribution, the corresponding number of the estimated $q(w \mid d)$ is lower than for LDA. Hence, the performance improvement is reasonable. In addition, we find that for $K = 10$, the perplexity results start to converge. We also observed that when the number of aspects for an event is larger than 10, there will be more repeated aspects. This is also consistent with the empirical characteristics of real-world events. Usually, for a single event, there are rarely more than 10 focus points. Therefore, we set $K = 10$ for the rest of the experiments.

4.3 Results Analysis

The event aspect words are responsible for capturing the content features, through which we can determine what points the mainstream news and Twitter users focus on. We obtain 10 event aspects for each event and select the top-10 words for each aspect. In Table 3, we list partial results of 3 aspects that best describe the differences in focal points between news and Twitter. We can derive the following conclusions:

1. Traditional news media tends to be more rational, impersonal, and serious than Twitter. Twitter users are often more emotional in expressing their personal feelings, e.g., praying in response to the Nice attack in France (A4 in Table 3), or for the Orlando shooting victims.

2. News media tends often considers the political background and implications. For instance, it reviews the previous Nobel Peace Prize winners (A3) and relates the gun control debate for the Orlando shooting with the elections.
3. The news media tend to be more comprehensive in fully covering relevant background and different aspects of an event. For instance, they provide relevant background knowledge to let the readers better understand the circumstances of an event, e.g., enumerating different countries relevant to "TPP" (A2). Twitter users care more about how the event may impact their personal life, e.g., users discuss the influence of the immigration ban policy on employment in US tech companies (A5).
4. News media shows a wider coverage of aspects, including important aspects that however may not directly affect a large percentage of their readership, e.g., the protest against the pipeline construction in Dakota (A1).
5. When discussing event-related topics, Twitter users focus more on major protagonists or celebrities. Twitter users tend to express their opinions about noteworthy people involved in the event (A6).

Table 3. Selected event aspect words in news and Twitter

News specific aspects			Twitter specific aspects		
A1	A2	A3	A4	A5	A6
Pipeline	Trade	ShimonPeres	Nice	Immigration	BobDylan
Dakota	USA	Prime	Terrorist	Ban	Literature
North	Countries	Minister	PrayForNice	Employees	Arrogant
Rock	Pacific	Israel	People	Google	Congrats
Standing	China	YitzhakRabin	Prayers	Tech	Impolite
Access	Deal	President	Victims	Christo	Winning
Army	Agreement	Peace	BastilleDay	Art	Called
Oil	Australia	Party	Sad	Comcast	Silence
Sioux	Pact	Leader	Mourning	World	Wind
Tribe	Zealand	YasserArafat	Families	Companies	Knock

Due to space constraints, we can't show the category words. However, we find that some words succeed in representing the important characteristics of a given category. For instance, *police*, *victims* in the attack category, *residents*, *damage* in the disaster category, *price*, *market* the in business category, etc.

5 Conclusion

In this paper, we have presented a new method EvA to analyze real-world events, combining news and social media. And we use this model to discover important

distinctions between mainstream news media and Twitter postings, based on the aspect analysis. In terms of future work, we will attempt to exploit our conclusions in several kinds of applications and follow-up studies. One direction is to consider classifications beyond the content-based category labels that we have considered thus far. Further kinds of classifications include those pertaining to the life cycle (short, middle, and long-term period), popularity (popular or not), authenticity (trustworthy or not), etc.

Acknowledgements. The authors wish to acknowledge the support provided by the National Natural Science Foundation of China (61503217, 91546203), the Key Research and Development Program of Shandong Province of China (2017CXGC0605) and China Scholarship Council (201606220187). Gerard de Melo's research is funded in part by ARO grant W911NF-17-C-0098 (DARPA SocialSim).

References

1. Abdelhaq, H., Sengstock, C., Gertz, M.: EvenTweet: online localized event detection from Twitter. Proc. VLDB Endow. **6**(12), 1326–1329 (2013)
2. Atefeh, F., Khreich, W.: A survey of techniques for event detection in Twitter. Comput. Intell. **31**(1), 132–164 (2015)
3. Castillo, C., El-Haddad, M., Pfeffer, J., Stempeck, M.: Characterizing the life cycle of online news stories using social media reactions. In: CSCW, pp. 211–223 (2014)
4. Hänska-Ahy, M., Wardle, C., Browne, M.: Social media & journalism: reporting the world through user generated content. Particip.: J. Audience Recept. Stud. **10**, 436–439 (2013)
5. Hoffart, J., Milchevski, D., Weikum, G.: STICS: searching with strings, things, and cats. In: SIGIR, pp. 1247–1248 (2014)
6. Hu, Y.: Event analytics on social media: challenges and solutions. Arizona State University (2014)
7. Hu, Y., Hong, Y.: Modeling Twitter engagement in real-world events. In: HICSS (2017)
8. Hu, Y., John, A., Seligmann, D.D., Wang, F.: What were the tweets about? Topical associations between public events and Twitter feeds. In: ICWSM (2012)
9. Hu, Y., John, A., Wang, F., Kambhampati, S.: ET-LDA: joint topic modeling for aligning events and their Twitter feedback. In: AAAI, vol. 12, pp. 59–65 (2012)
10. Li, C., Bendersky, M., Garg, V., Ravi, S.: Related event discovery. In: WSDM, pp. 355–364. ACM (2017)
11. Olteanu, A., Castillo, C., Diakopoulos, N., Aberer, K.: Comparing events coverage in online news and social media: the case of climate change. In: ICWSM, No. EPFL-CONF-211214 (2015)
12. Paul, M., Girju, R.: Cross-cultural analysis of blogs and forums with mixed-collection topic models. In: EMNLP, pp. 1408–1417. ACL (2009)
13. Setty, V., Anand, A., Mishra, A., Anand, A.: Modeling event importance for ranking daily news events. In: WSDM, pp. 231–240. ACM (2017)
14. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In: SIGCHI, pp. 1079–1088. ACM (2010)
15. Zhao, W.X., et al.: Comparing Twitter and traditional media using topic models. In: Clough, P., et al. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-5_34