Multifaceted Event Analysis on Cross-Media Network Data

Daheng Wang, Meng Jiang, Xueying Wang, Nitesh Chawla, Paul Brunts
University of Notre Dame, Notre Dame, Indiana, 46556, USA
{dwang8, mjiang2, xwang41, nchawla, pbrunts}@nd.edu

ABSTRACT

People are discussing about events such as shootings, protests, flight crashes and new public policies across news media and social media. Different media sources reflect different facets of the events. In this positioning paper, we point out the importance of integrating various data from the multiple media sources for understanding/analyzing the events. Here the big data's "Variety" issue can be resolved by extending the information network representation to a "cross-media information network" but how to generate comprehensive event analysis from the network's components? We propose a multifaceted analysis framework that models four critical facets to fully understand the events across social media network and news media network. The facets include *media type* for differentiating event information sources; *content* for discovering events from unstructured text; *sentiment* on quantifying user reflecting on events; and, *time* for turning events into dynamic evolving objects. We performed our framework on a real data set from Twitter and Google News, which creates interesting and useful event description and visualization for human inspection.

KEYWORDS

multifaceted analysis, event analysis, cross-media network, heterogeneous network

ACM Reference format:
Daheng Wang, Meng Jiang, Xueying Wang, Nitesh Chawla, Paul Brunts University of Notre Dame, Notre Dame, Indiana, 46556, USA {dwang8, mjiang2, xwang41, nchawla, pbrunts}@nd.edu . 2018. Multifaceted Event Analysis on Cross-Media Network Data. In *Proceedings of International Workshop on Heterogeneous Networks Analysis and Mining, Los Angeles, California, USA, February 2018 (HeteroNAM'18),* 8 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

With the development of Internet, people are getting closer during all kinds of real life events by utilizing usually more than one type of media platforms. Different media platforms are designed to focus on different aspects of the events. For example, traditional media platforms such as CNN, Reuters, and CNBC, have comprehensive and up-to-date coverage on emergency events; while social media platforms such as Facebook, Twitter, Instagram, allow the public to engage in discussion along the development of events, which largely facilitates information propagation. Information on each media platform can be considered as a heterogeneous information network containing entities of users, information products and relations between them. Though each information network is huge itself, it acts as one component of a giant cross-media information network, in which different components are connected by events. We claim that studying the interaction between different media network components provide us valuable and deep insights into understanding and characterizing real life events, which cannot be discovered in any single media network.

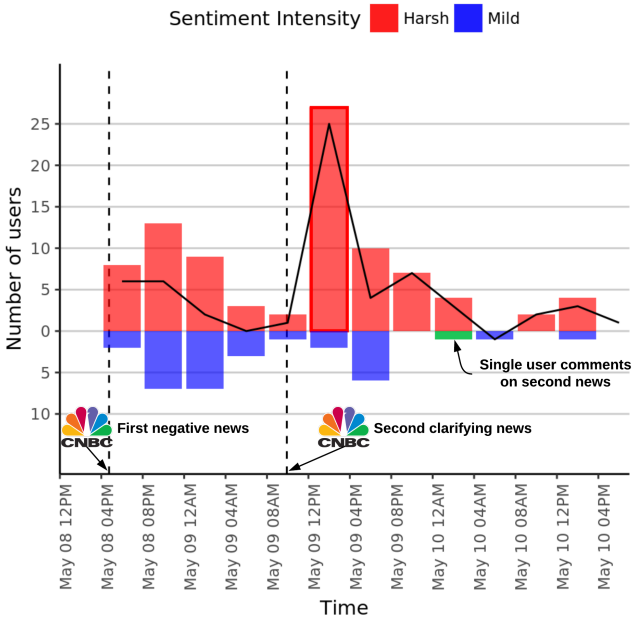

Figure 1: Dynamics of Twitter user sentiment on "*contrast_comments*" event.

Motivation. One special event happened lately, which is quite illustrative in this scenario, we denote it as the "*contrast_comments*" event in this paper. The brief storyline of this event is as follows: on someday in May 2017, CNBC[1] published a news article on its website which contains criticism and reckless negative comments from a venture capitalist on a technology company. This quickly drew a large amount of attention from users interested in technology and stock market on Twitter[2]. On the second day morning, another news article reporting the same venture capitalist came up with directly opposite attitude on the same technology company, which is a follow-up clarification message, appeared on CNBC website. But unlike the first time, there were fewer discussions and reactions from users on Twitter to comment on this specific follow-up message. We were able to collect the post stream from Twitter on

HeteroNAM'18, February 2018, Los Angeles, California, USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

[1] https://www.cnbc.com/
[2] https://twitter.com/

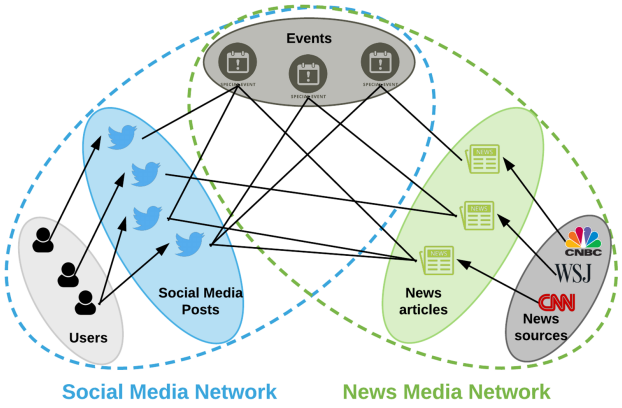

Figure 2: Cross-media information network.

this event for 3 days starting from the publishing time of the first news. To better understand people's refection, we also read all the tweets and tagged them into one of 3 levels of sentiment intensity: *neutral*, *mild*, or *harsh*. Then, we visualized the development of "*contrast_comments*" in Figure 1.

In the figure, a general impression we have is that much more users have harsh sentiment on this event compared with mild sentiment. From left to right, we can see the difference of harsh sentiment and mild sentiment (indicated by the black solid line) starts high at the publishing time of first news, gradually declines till the end of the day, and then have a huge pike right after the publishing time of the second news. When we take a closer look at the huge pike (emphasized by red bold border), we find it is interesting and counter-intuitive: (1) our intuition would say that the second clarifying news would change the sentiment of users on this event later on, and (2) the second clarifying news would transfer users' attention from the first news into the new one. However, it is easy to see from the Figure 1 that people inclined to have harsh sentiment after the second news, and almost nobody talks about the second news (only one highlighted by green color).

A few important observations we can make from the example of "*contrast_comments*" event include: (1) the environments of developing event are complex and interactive. News reports from news media can influence the sentiment of users on social media, and in turn, the unusual sentiment of users on social media can also cast pressures on news media and brew changes; (2) the development of event in such complex environments is also complicated and dynamic, mixed with various types of stimuli or obstacles. Therefore, it is straightforward to ask "*how can we depict the complex environments?*", and "*how can we characterize and understand the development of event in a given environment?*" for event analysis.

The first question can be answered by modeling the complex environment as a cross-media network, like the one shown in Figure 2. A social media network and a news media network are the two major components of cross-media network, overlapping on their common latent event entities. Just like social media network and news media network, such a composite network is also a heterogeneous network. Then, the crux question becomes "*how to study the dynamics of event on the cross-media network*".

In this paper, we adopt a data driven approach to address these issues. We propose a multifaceted event analysis framework for studying dynamics of event development across different types of media platforms. In particular, motivated by our previous example, four distinct and indispensable event analysis facets for capturing the dynamics of event are bundled into this framework: *media type*, *content*, *sentiment*, and *time*. The media type facet lies down the foundation for the other three facets toward cross-media property; the content facet allows us to discover latent common events exist on cross-media network; then, the sentiment facet provides us a quantitative view of human feelings and reactions on event; last but not least, the time facet gives us the ability to transform event from static object into evolving entities.

To test the effectiveness of our multifaceted event analysis framework, we built a real-world cross-media network based on data sets we collected from Twitter and Google News for 4 months. Instead of aiming at training an event prediction model, we apply our framework to generate interesting and useful event description and visualization for human inspection, which can provide more valuable insights into understanding event on cross-media network.

As listed below, the major contributions of this work can be summarized as below.

(1) We formally define cross-media network as as composite heterogeneous network of social media network and news media network. Furthermore, we give out a multifaceted event analysis framework on cross-media networks.
(2) We illustrate useful and interesting findings produced by applying our multifaceted event analysis framework on a real-world cross-media network data set.
(3) Also, we provide a useful tool[3] specially prepared to helps readers in performing the aforementioned analysis task on other cross-media data sets.

The roadmap of this paper is as follows: we review basic concepts about information network in Section 2; in Section 3, we give formal definition on cross-media network and present our multifaceted event analysis framework; Then, we show our experimental results in Section 4; related work will be discussed in Section 5; finally, we conclude this paper in Section 6.

2 BASIC CONCEPTS AND DEFINITIONS

In this section, we review basic concepts about heterogeneous information network. We adopt the definition given by Sun et al. [18]. Then we give examples of modeling *social media information network* ("social media network" for short) and *news media information network* ("news media network" for short) as two typical heterogeneous information networks.

2.1 Heterogeneous Information Network

A heterogeneous information network is a special type of information network with the underneath data structure as a directed graph, which either contains multiple types of objects or multiple types of links.

Definition 2.1. **Information Network** [18]. An information network is defined as a directed graph $G = (V, E)$ with an object type mapping function $\phi : V \rightarrow \mathcal{A}$ and a link type mapping function $\psi : E \rightarrow \mathcal{R}$, where each object $v \in V$ belongs to one particular

[3] https://github.com/adamwang0705/cross_media_affect_analysis

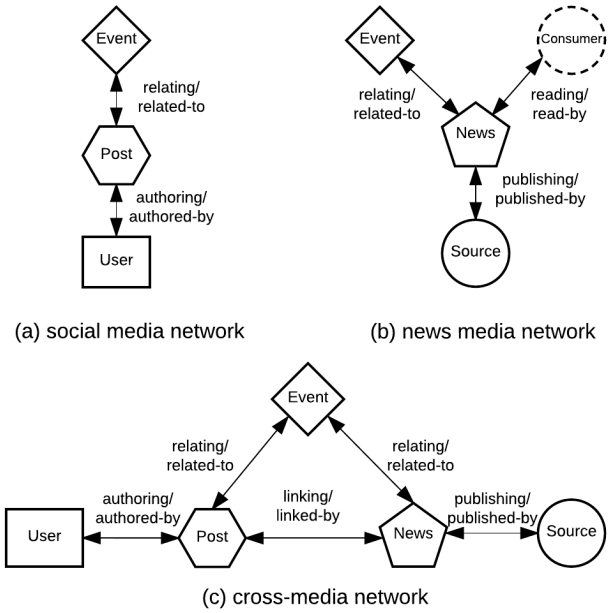

Figure 3: Network schema of social media, news media, and cross-media networks.

object type $\phi(v) \in \mathcal{A}$, and each link $e \in E$ belongs to a particular relation $\psi(e) \in \mathcal{R}$. If two links belong to the same relation type, they share the same starting object type and the ending object type.

Different from the traditional network definition, we explicitly distinguish object types and relationship types in the network, which means the types of objects $| \mathcal{A} | > 1$ or the types of relations $| \mathcal{R} | > 1$, the network is called **Heterogeneous Information Network** ("heterogeneous network" for short); otherwise, it is a homogeneous information network.

To better understand the object types and link types in a complex heterogeneous network, it is necessary to provide the meta level (i.e., schema-level) description of the network. The following concept of network schema is defined to describe the structure of a network.

Definition 2.2. **Network schema** [18]. The network schema is a meta template for a heterogeneous network $G = (V, E)$ with the object type mapping $\phi : V \rightarrow \mathcal{A}$ and the link mapping $\psi : E \rightarrow \mathcal{R}$, which is a directed graph defined over object types $\mathcal{A}$, with edges as relations from $\mathcal{R}$, denoted as $T_G = (\mathcal{A}, \mathcal{R})$.

The network schema of a heterogeneous information network specifies type constraints on the sets of objects and relations. These constraints make a heterogeneous network semi-structured, guiding the semantics explorations of the network [16].

2.2 Social Media and News Media Networks

Heterogeneous network is ubiquitous in modern information society. Specifically, we can model information from two types of online media platforms as heterogeneous network: one is social media network, and the other one is news media network.

Example 2.3. An **Social Media Network**, e.g., Facebook or Twitter, is a typical heterogeneous network, as shown on the left side of Figure 2. Entity types of user (U) and web post (P, e.g. article on Facebook and tweet on Twitter), lie in the heart of this heterogeneous social network. Related to the web post, event (E) serves as the type of objects depicting the natural and meaningful real-life incidents appear in posts. Each post $p \in P$ is linked to a user and a set of events. Link types are defined by these relations respectively.

Note that a social media network may also include other types of entities such as hashtag, geographical location, photo tag, or video playlist, depending on the specific setting of the platform. We focus on the subgraph constituted by objects of users, posts, events, and relations between them in this study.

Example 2.4. A **News Media Network** can be modeled as heterogeneous network, as shown on the right side of Figure 2. News article (N) and news source agency (S) are the central object types in this heterogeneous news network. Entity type of event (E) also plays an important role as in the social media network. Furthermore, we have the object type of news consumer (C) for the audience of news articles. Each $n \in N$ is linked to a source agency, a set of events, a set of consumers, and their link types are defined by these relations respectively.

Other possible object types in a news media network may contain journalist, editor, category, or comment. In this work, we constraint ourselves to study the subgraph made up by objects of news articles, sources, consumers, events, and the relations between them.

Example 2.5. **Social Media Network Schema** and **News Media Network Schema**. Figure 3 (a) shows the network schema for social media network that was defined in Example 2.3. Figure 3 (b) shows the network schema of news media network that was defined in Example 2.4. In the social media network schema, links between posts and users denoting the authoring/authored-by relations; between posts and events denoting the relating/related-to relations. Similarly, in the news media network schema, links between sources and news denoting publishing/published-by relations; between events and news for relating/related-to relations; and, between news and consumers for reading/read-by relations.

3 CROSS-MEDIA MULTIFACETED EVENT ANALYSIS

In this section, we introduce our methodology of evaluating events on cross-media network data. First, we define cross-media network as a fusion of social media network and news media network, hinged upon a pivotal type of object, i.e., event that is worth the attention of both social media users and news media readers. The real-life events naturally have multiple facets of information generated from different information sources. We propose a novel multifaceted event analysis framework on cross-media network.

3.1 Cross-Media Information Network

Based on social media network and news media network we have explained in Example 2.3 and Example 2.4, we define the main subject of this study: cross-media information network as a composite heterogeneous network of these two network parts.

Definition 3.1. A **Cross-Media Information Network** (cross-media network for short) is a heterogeneous network consisting

two key components: (1) a heterogeneous social media network, and (2) a heterogeneous news media network. There are 5 object types in a cross-media network: user (U), post (P), event (E), news (N), and source (S). The former 3 entity types comes from the social media network side, and the latter 3 exist in the news media network. Two major components of cross-media network overlap on their common event objects and interleave with each other between objects of event, post, and news. Link types are defined by corresponding relations between objects.

The network schema of cross-media network and detailed relations between objects are shown in Figure 3 (c). Note that, slightly different from what we defined in Example 2.4, there are no consumer object type in a cross-media network. Since we are primarily interested in the news audience who are also observable on social media network, the consumer object type in news media network collapses into the user object type on social media network side.

3.2 Multifaceted Event Analysis Framework

In this work, we take a data-driven approach and propose the framework of multifaceted event analysis on cross-media network data to systematically characterize real-life events across media networks. Specifically, we inflate our event analysis space along 4 different axes: **media type**, **content**, **sentiment**, and **time**. Each one of these facets in our event analysis framework is different to each other. Yet, as we are going to explain below, they are all substantive and indispensable for any study of events on cross-media network.

*3.2.1 the **media type** facet.* First, the facet of **media type**, serving as a cornerstone, lies down the foundation for the other 3 facets in our multifaceted event analysis framework. It generally refers to different information origins of event in the cross-media network. In Definition 3.1, we have specified cross-media network as a composite heterogeneous network of social media network and news media network. We say that the *media type* of events on such a cross-media network links to two origins: the origin of social media, and the origin of news media. In this work, by taking a data-driven perspective, we focus on two specific media types of event: Twitter social network, and online public news media represented by aggregation of news in Google News[4]. Nevertheless, as the definition of cross-media network can be easily extended to include more than one social/news media network, this facet of event analysis is also inherently capable of capturing multiple media types of event information.

*3.2.2 the **content** facet.* Secondly, the **content** facet enables us to discover events scattered on cross-media network as latent entities. We rely on the text content of social media posts and news articles to distill out information objects representing real-life events. Previous work utilize various type of language models to learn events happening either on social media network or news media network [2, 10, 14]. In our study, since we are provided with succinct news title data, which is usually in high quality and produced by professional editors, we took the naive approach and chose to manually select out salient and popular real-life events among these news articles. High frequency terms in news titles were extracted during the process to help us accomplish this task. Furthermore, we labeled each event into one of 4 broad event categories: politics, social, entertainment, or tragedy. At last, we have 51 events spread almost evenly across four categories. In Table 1, we show some representative events from each category. The complete information about all events can be found in Appendix A.

[4]https://news.google.com/

Table 1: Representative events in each category

Category	Name	#news
politics	Hillary Clinton's email controversy	228
politics	Ukraine cease fire negotiation	84
social	Sony under cyberattack	275
social	Patriots Deflategate	44
entertainment	57th Annual Grammy Awards	99
entertainment	Christmas holiday	237
tragedy	AirAsia Flight QZ8501 crash	258
tragedy	2014 FSU shooting	39
...	... ...	...

*3.2.3 the **sentiment** facet.* Thirdly, the **sentiment** facet provides us a quantitative view of human feelings and reactions on event, comparable between different information origins or event categories. The variance of sentiment across media types can provide us additional insights into event that cannot be acquired in a single media type network. Generally speaking, sentiment refers to the subjective mental state of an individual. In this work, we are interested in the people's collective sentiment inclination and patterns displayed on cross-media network.

There exist a dozen of techniques to measure human sentiment in online environment. Conventional supervised methods train a classifier on labeled training data and apply the model on test data to predict the sentiment value of unknown text data [6, 9]. In the example of "*contrast_comments*" event given in Section 1, we asked human judges to rate each tweet into one of 3 scales of sentiment intensity because of the pervasiveness of sarcastic tone, which is hard for machine learning algorithms to detect, in related tweets. Since we are not provided with any sentiment label on our cross-media network data, to automate the rating of sentiment and scale it to the whole network level, we adopt the Hedonometer metric proposed by Dodds et al. [3].

Hedonometer measures the happiness level of texts that hinges on two key components: (1) human evaluations of the happiness of a set of individual words, and (2) a naive algorithm for scaling up from the individual words to texts. Specifically, for a given text T, the weighted average level of happiness for the text is computed as

$$h_{avg}(T) = \frac{\sum_{i=1}^{N} h_{avg}(w_i) f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{avg}(w_i) p_i \qquad (1)$$

where f_i is the frequency of the ith word w_i for which we have an estimate of average happiness, $h_{avg}(w_i)$, and $p_i = f_i / \sum_{i=1}^{N} f_i$ is the corresponding normalized frequency. A list of 10,222 most frequent words drew on a large corpus, including Twitter, news media reports, and books, were evaluated by users on Amazon Mechanical

Turk[5]. Each word got 50 independent evaluations on a nine point integer scale ranging from 1 to 9 (larger value for being happier). A few examples are $h_{avg}(laughter) = 8.50$, $h_{avg}(reunion) = 6,96$, and $h_{avg}(greed) = 3.06$. We also follow the authors' recommendation to exclude words lies in $5 - \Delta h_{avg} < h_{avg} < 5 + \Delta h_{avg}$ with $\Delta h_{avg} = 1$, which results in a subset of 3,686 sentimental words that can be quantified.

Besides using a single summary statistic such as Hedonomter to represent human sentiment on events, it is also naturally to analyze the difference of sentiments lie between events by taking a deeper look at the changes of underlying word frequency [3]. In particular, for a reference text T_{ref} and a comparison text T_{comp} with happiness scores $h_{avg}^{(ref)}$ ang $h_{avg}^{(comp)}$, we can write

$$h_{avg}^{(comp)} - h_{avg}^{(ref)} = \sum_{i=1}^{N} h_{avg}(w_i)[p_i^{(comp)} - p_i^{(ref)}] \quad (2)$$

combining with a simple transformation

$$\sum_{i=1}^{N} h_{avg}^{(ref)}[p_i^{(comp)} - p_i^{(ref)}] = h_{avg}^{(ref)})(1-1) = 0$$

, we can rewrite Equation (2) as

$$h_{avg}^{(comp)} - h_{avg}^{(ref)} = \sum_{i=1}^{N} \underbrace{[h_{avg}(w_i) - h_{avg}^{(ref)}]}_{+/-} \underbrace{[p_i^{(comp)} - p_i^{(ref)}]}_{\uparrow/\downarrow} \quad (3)$$

Then, the contribution of the ith word to the difference $h_{avg}^{(ref)} - h_{avg}^{(comp)}$ can be characterized from two aspects:

(1) whether or not the ith word is on average happier than $h_{avg}^{(ref)}$ (denoted by $+/-$), and
(2) whether or not the ith word appears more often in T_{comp} than in T_{ref} (denoted by $\uparrow$ /$\downarrow$)

*3.2.4 the **time** facet.* Last but not least, the **time** facet helps us to transform events from static objects into dynamic evolving entities. As illustrated by the "*contrast_comments*" example in Section 1, any real-life event would be continuously changing, as well as the sentiment of the public toward it. To get a proper and in-depth understanding of event dynamics on a cross-media network, temporal information accompanied with **content** and **sentiment** from both **media types**, i.e., social media and news media, becomes essential and necessary.

4 EXPERIMENTS

We explain our data collection mechanism here and summarize the cross-media network data set we have built. Then, we present 3 interesting and useful findings on this cross-media network data set by focusing on different subset of analysis facets from our multifaceted event analysis framework.

4.1 Data set

As mentioned in Section 3.2.1, we built up our cross-media network data set by carefully aligning data from Twitter and Google News. Direct method of collecting data from both sources in parallel does not work in our case for two reasons: (1) the public sample tweets returned from Twitter Sample real-time Tweets API[6] are in tremendous amount, most of which are irrelevant to the news data we extracted from Google News; and, (2) we also need additional steps to establish relations between tweets and news articles.

Therefore, to simply the process of building up cross-media network and maintaining a high quality in the meantime, we used the following data collection mechanism: An automatic crawler was deployed to scrape news articles from Top Stories section of Google News on half-hour basis. For each news article returned by the crawler, we extracted non phrases from its title and snippet on-the-fly and pass them to the Twitter Standard search API[7] for querying. Tweets containing at least two non phrases were accepted. And, we collected tweets timestamped within one day after the publishing time of the news article.

We have experienced Twitter API upgrade and occasional server errors during the data collection process. So, only a subset of data collected between Nov. 18 2014 and Apr. 14 2014 were taken for consistency[8]. Within this date range, we have 37,286 news articles with 176,288,149 related tweets in total. We further filtered out news or tweets unrelated to any one of the 51 events we manually picked (specified in Section 3.2.2). At last, we have 6,125 news articles published by 569 different news sources, associated with 37,608,586 tweets authored by 19,471,517 unique users. These objects and corresponding relations made up the cross-media network data set we used in following experiments.

4.2 Sentiment Differs Across Media Types and Categories

To get an overview of our cross-media network data set, we first explore the leading three analysis facets: **media type**, **content**, and **sentiment**. For each event e, we aggregate all news related to it into a document $D_e^{(n)}$, and all tweets related to it into $D_e^{(p)}$. Then, a single average happiness score is calculated on both $D_e^{(n)}$ and $D_e^{(p)}$ according to Equation 1. The distribution of event happiness scores across media types and categories is shown in Figure 4.

A general impression when looking at the figure is that: almost all distribution boxplots lie well above the neutral line, i.e., value 5 on y-axis, regardless of media types or categories, though a few exceptions can be found on the social media side in the category of politics and tragedy. This is consistent with the observation made by the authors of our adopted Hedonometer metric: there exists a positive trend in the language usage of general English [3].

Besides, it is easy to notice that the distribution of happiness scores calculated on $D_e^{(p)}$ in all categories have a higher median value than happiness scores on $D_e^{(n)}$. In other words, on the same event, people generally display more optimistic mood on social media when compared with news media, via the representation of their written expressions. This is especially salient in the event

[5]https://www.mturk.com/mturk/welcome

[6]https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_statuse_sample

[7]https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

[8]Though we try to minimize inconsistency, data on these dates are unavailable due to miscellaneous reasons: Dec. 15 2014, Mar. 23 to Mar. 28 2015

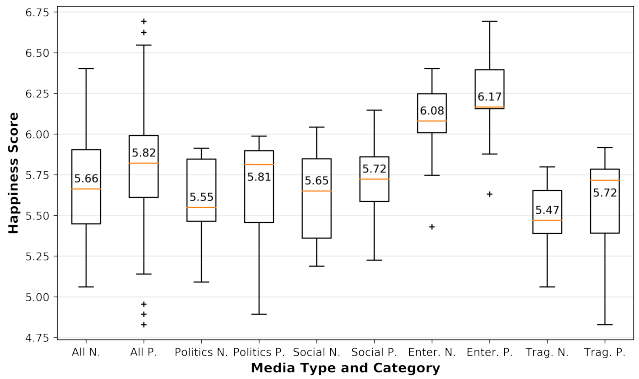

Figure 4: Distribution of sentiment on events across media types and categories (N. for news and P. for tweets).

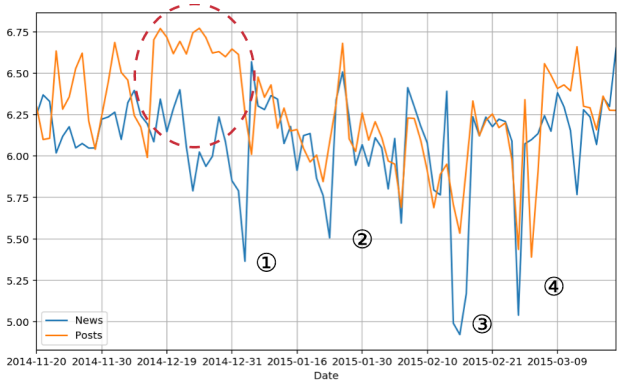

Figure 5: Sentiment dynamics of entertainment events

category of entertainment, where all quantile values are higher on social media side.

We further check the Spearman correlation of happiness scores between two media types to see, for the same event, whether the general sentiment trend on news media indicates similar trend on social media, or vice versa. The overall correlation coefficient is 0.73 without consideration of event categories, which is a fairly high value. However, when we break it down into different categories, things become interesting. The correlation coefficient is 0.34 and 0.11 for category of social and entertainment; 0.70 and 0.60 for category of politics and tragedy. It is not hard to understand that people have similar less happy feelings on tragedy events, and more often, on controversial political affairs. But clearly, people display very diverse and unrelated sentiments on social life and entertainment events across two different media types.

4.3 Entertainment Events Sentiment Dynamics

After we have acquired an overall understanding of our cross-media network, we now extend our analysis space to include the **time** facet from our multifaceted event analysis framework. In particular, we expand the **sentiment** dynamics along **time** axis across **media types**, focusing on the **content** in category of entertainment. Given a time period t, say a day, we assemble the document $D_t^{(n)}$ for all news related to any entertainment event, and $D_t^{(p)}$ for all tweets in similar way. Happiness scores are calculated on daily news and tweets documents, as shown in Figure 5.

We can see that, in the figure, the orange line indicating tweets sentiment roughly follows the blue line for news sentiment. There is a large deviation of two lines starting from Dec. 18 2014 to Jan. 3 2015 (marked by the red dash circle). This gap can be well explained by events such as *Thanksgiving*, *Black_Friday_and_Cyber_Monday*, and *Christmas* (31, 32, 33, in Table 2). During these two weeks, there is a upsurge of joyful moods on social media side, but the change is softer in news media. After that, we can identify a small lag between the sentient trends of two media types, which indicates the sentiment trend on news media takes time to pass onto social media and exerts its influence on users.

We also notice a few drastic drops and annotated them with numbers in the figure. For the last three drops, we manually checked news and tweets on that date, and found they all attribute to a single event: *success_of_American_Sniper* (38 in Table 2). This event is related to the biographical war drama film American Sniper[9] released in 2014. It gets labeled as an entertainment event because most news and tweets talk about the success of this movie in box office. But this event is also mixed with news reports on the murder trial of the original book's author[10]. Whenever there is a new progress on the trail, it pulls down the line noticeably. However, the duality of *success_of_American_Sniper* does not explain the first drop occurs around Jan. 2 2015. This leads us to the next section of our experiments–a case study on the *2015_New_Year* event.

4.4 Case Study: Event of *2015_New_Year*

At last, to further refine our observations made in section 4.3, we present a case study on the event of *2015_New_Year*. Hoping to highlight and explain the odd uniqueness of sentiment showing on Jan. 2 2015, we build a reference document $D_{t'}^{(ref)}$ of news, where Dec. 18 2014 $\leq t' \leq$ Jan. 1 2015. The news associated with *2015_New_Year* are used to build the comparison document $D_{e'}^{(comp)}$, where e' denotes the event of *2015_New_Year*. Then, we employ Equation 3 to find out words contributing most to the confusing "bad mood on new year".

Surprisingly, words appear more often in $D_{e'}^{(comp)}$ with lower happiness values, i.e. $(-\uparrow)$ words, include "killed", "tragedy", "injured", "victims", and "accident". Other words with abnormal high frequency include "Shanghai" and "stampede". We turned back to examine all the news articles related to the *2015_New_Year* event and found that: 21 out of 69 of them are actually reporting a mass stampede tragedy happened on the 2015 New Year's eve in Shanghai[11]. This may not be a wonderful discovery, but it still illustrates another possibility of defining and analyzing event in finer granularity.

[9] https://en.wikipedia.org/wiki/American_Sniper
[10] https://en.wikipedia.org/wiki/Chris_Kyle
[11] https://en.wikipedia.org/wiki/2014_Shanghai_stampede

5 RELATED WORK

News content features and temporal features have been well studied in the domain of Retrospective news Event Detection (RED) [2]. Earlier work of event detection and analysis on social network also relies on the text content of social media posts [14]. Later, researchers start to leverage the geo-location and temporal information commonly accompanied with social media posts [10, 13]. Recently, the topological information of social media network have also been found helpful for analyzing events on social media [1]. Although these work try to utilize various types of information when analyzing event, they do not consider the interactions between different types of media networks. In other words, the scope of these studies is limited to a single type of media network. Our approach of event analysis treats social and news media network as components of the global cross-media network and leverage common events to bridge them together.

There are early studies jointly work on social media and news media [5, 7, 8, 11, 17]. Sun et al. [18] first proposed the definition of heterogeneous network. After that, a large number of applications have been built on top of it, such as link prediction [19], representation learning [4], and fact checking [15]. Some recent studies are related to this work that they also deal with the heterogeneity in cross-media network [12, 20]. However, previous work primarily focus on measuring the correlation between contents from different media networks. Our work is fundamentally different from them in the way that we propose a multifaceted analysis framework to systematically evaluate events.

6 CONCLUSIONS

In this paper, we studied the characteristics of event on cross-media network from multiple angles. We modeled cross-media network as a composite heterogeneous network of social media and news media. And, we proposed a multifaceted event analysis framework which incorporate facets of media type, content, sentiment, and time. We applied this framework on a real cross-media network data set and contributed some interesting and useful discoveries.

There are a few directions we can take next: first, we just made a first step toward a comprehensive event analysis. Other useful facets, such as geo-location, can be added into our analysis framework. One useful application can be building an automatic event classifier by incorporating additional facets or features from cross-media network; secondly, an automatic method for identifying latent semantic objects in social and media network can be greatly beneficial for extracting event objects among them. This can help us handle larger data sets and analyze more events, leading us to go beyond the proof-of-concept type of experiments; thirdly, it would also be helpful to test on building cross-media network covering more than 2 different media types, which may yield more valuable insights into event analysis across multiple type of media platforms.

A THE SELECTED EVENTS FOR MULTIFACETED CROSS-MEDIA ANALYSIS

Here, we give the detailed information of all 51 manually selected and inspected events in Table 2.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions.

REFERENCES

[1] Shankar Bhamidi, J Michael Steele, Tauhid Zaman, and Others. 2015. Twitter event networks and the superstar model. *The Annals of Applied Probability* 25, 5 (2015), 2462–2502.
[2] Xiangying Dai, Yancheng He, and Yunlian Sun. 2010. A Two-layer Text Clustering Approach for Retrospective News Event Detection. In *Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference on*, Vol. 1. IEEE, 364–368.
[3] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one* 6, 12 (2011), e26752.
[4] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 135–144.
[5] Wei Gao, Peng Li, and Kareem Darwish. 2012. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1173–1182.
[6] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1, 2009 (2009), 12.
[7] Meng Jiang, Peng Cui, Nicholas Jing Yuan, Xing Xie, and Shiqiang Yang. 2016. Little Is Much: Bridging Cross-Platform Behaviors through Overlapped Crowds.. In *AAAI*. 13–19.
[8] Meng Jiang, Christos Faloutsos, and Jiawei Han. 2016. Catchtartan: Representing and summarizing dynamic multicontextual behaviors. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 945–954.
[9] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! *Icwsm* 11, 538-541 (2011), 164.
[10] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. 2012. Tedas: A twitter-based event detection and analysis system. In *Data engineering (icde), 2012 ieee 28th international conference on*. IEEE, 1273–1276.
[11] Lu Liu, Feida Zhu, Meng Jiang, Jiawei Han, Lifeng Sun, and Shiqiang Yang. 2012. Mining diversity on social media networks. *Multimedia Tools and Applications* 56, 1 (2012), 179–205.
[12] Xiaozhong Liu, Tian Xia, Yingying Yu, Chun Guo, and Yizhou Sun. 2016. Cross Social Media Recommendation.. In *ICWSM*. 221–230.
[13] Fred Morstatter, Shamanth Kumar, Huan Liu, and Ross Maciejewski. 2013. Understanding twitter data with tweetxplorer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1482–1485.
[14] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.
[15] Baoxu Shi and Tim Weninger. 2016. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 101–102.
[16] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2017. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2017), 17–37.
[17] Ilija Subašic and Bettina Berendt. 2011. Peddling or creating? investigating the role of twitter in news reporting. *Advances in Information Retrieval* (2011), 207–213.
[18] Yizhou Sun and Jiawei Han. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* 3, 2 (2012), 1–159.
[19] Yang Yang, Nitesh Chawla, Yizhou Sun, and Jiawei Hani. 2012. Predicting links in multi-relational and heterogeneous networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 755–764.
[20] Yang Yang, Yizhou Sun, Jie Tang, Bo Ma, and Juanzi Li. 2015. Entity matching across heterogeneous sources. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1395–1404.

Table 2: Manually select events

Id	Category	Name	Span days (start/end)	#news	#tweets
1	politics	Hillary_Clinton_email_controversy	38 (2015-03-02/2015-04-09)	228	367,618
2	politics	Iran_nuclear_deal	146 (2014-11-19/2015-04-14)	406	1,238,107
3	politics	ISIS_Jihadi_John_identity_reveal	35 (2015-02-01/2015-03-08)	101	310,280
4	politics	Ukraine_cease_fire	124 (2014-12-02/2015-04-05)	84	233,999
5	politics	Egypt_free_Al_Jazeera_journalist	55 (2015-01-01/2015-02-25)	50	39,845
6	politics	Keystone_XL_Pipeline_bill	104 (2014-11-18/2015-03-02)	55	37,041
7	politics	CIA_Torture_Report	17 (2014-12-05/2014-12-22)	41	84,081
8	politics	Obama_cybersecurity_plan	110 (2014-12-19/2015-04-08)	73	249,609
9	politics	DHS_funding_issue	28 (2015-02-03/2015-03-03)	45	66,855
10	politics	US_Cuba_relationship	132 (2014-12-03/2015-04-14)	235	746,329
11	politics	2015_CPAC	4 (2015-02-25/2015-03-01)	68	111,755
12	politics	Iraq_free_ISIS_Tikrit	36 (2015-03-02/2015-04-07)	94	224,805
13	politics	Nigeria_Boko_Haram_terrorists	137 (2014-11-28/2015-04-14)	243	187,452
14	social	Ferguson_unrest	147 (2014-11-18/2015-04-14)	611	1,616,426
15	social	Hong_Kong_protest	135 (2014-11-18/2015-04-02)	157	110,839
16	social	Sony_cyberattack	97 (2014-11-25/2015-03-02)	275	902,546
17	social	Bill_Cosby_sexual_assault_allegation	93 (2014-11-18/2015-02-19)	168	241,487
18	social	SpaceX_rocket_landing	99 (2015-01-05/2015-04-14)	86	159,027
19	social	Brian_Williams_fake_story	79 (2014-12-03/2015-02-20)	69	131,549
20	social	HSBC_tax_scandal	10 (2015-02-08/2015-02-18)	28	39,947
21	social	David_Carr_death	3 (2015-02-12/2015-02-15)	36	23,416
22	social	Patriots_Deflategate	30 (2015-01-19/2015-02-18)	44	159,463
23	social	Delhi_Uber_driver_rape	54 (2014-12-07/2015-01-30)	36	199,832
24	social	Superbug_spread	57 (2015-01-07/2015-03-05)	41	159,846
25	social	Rudy_Giuliani_Obama_critique	104 (2014-12-07/2015-03-21)	50	195,681
26	entertainment	Oscar	113 (2014-11-20/2015-03-13)	241	993,397
27	entertainment	Super_Bowl	72 (2014-11-23/2015-02-03)	211	947,507
28	entertainment	Grammy	98 (2014-11-21/2015-02-27)	99	380,804
29	entertainment	Golden_Globe	32 (2014-12-11/2015-01-12)	79	413,222
30	entertainment	500_million_Powerball	16 (2015-02-07/2015-02-23)	79	246,179
31	entertainment	Thanksgiving	14 (2014-11-20/2014-12-04)	150	1,402,625
32	entertainment	Black_Friday_and_Cyber_Monday	13 (2014-11-21/2014-12-04)	121	995,610
33	entertainment	Christmas	86 (2014-11-28/2015-02-22)	237	2,123,840
34	entertainment	New_Year	85 (2014-12-25/2015-03-20)	69	698,497
35	entertainment	Apple_Watch	54 (2015-02-18/2015-04-13)	73	260,718
36	entertainment	Yosemite_historic_climb	21 (2015-01-08/2015-01-29)	41	23,084
37	entertainment	Jon_Stewart_Daily_Show	110 (2014-12-10/2015-03-30)	35	66,622
38	entertainment	success_of_American_Sniper	106 (2014-12-24/2015-04-09)	155	402,621
39	tragedy	Ebola_virus_spread	145 (2014-11-18/2015-04-12)	173	453,159
40	tragedy	Indonesia_AirAsia_Flight_QZ8501_crash	35 (2014-12-27/2015-01-31)	258	454,324
41	tragedy	Paris_attacks	120 (2014-12-04/2015-04-03)	225	684,566
42	tragedy	Vanuatu_Cyclone_Pam	6 (2015-03-13/2015-03-19)	89	81,207
43	tragedy	Malaysia_Airlines_Flight_MH370_crash	104 (2014-11-25/2015-03-09)	58	237,927
44	tragedy	Colorado_NAACP_bombing	110 (2014-11-29/2015-03-19)	38	128,994
45	tragedy	FSU_shooting	48 (2014-11-20/2015-01-07)	39	106,232
46	tragedy	Chapel_Hill_shooting	54 (2015-02-11/2015-04-06)	37	38,829
47	tragedy	Bobbi_Kristina_Brown_death	49 (2015-01-31/2015-03-21)	49	54,124
48	tragedy	Taliban_Pakistan_school_massacre	67 (2014-12-16/2015-02-21)	80	114,761
49	tragedy	American_ISIS_Hostage_Kayla_Mueller	34 (2015-02-07/2015-03-13)	38	21,529
50	tragedy	TransAsia_Airways_Flight_GE235_crash	4 (2015-02-03/2015-02-07)	56	77,302
51	tragedy	Germanwings_Flight_9525_crash	14 (2015-03-29/2015-04-12)	71	225,951