

Facebook, Twitter and Google Plus for Breaking News: Is there a winner?

Miles Osborne

School of Informatics
University of Edinburgh
EH8 9AB, Edinburgh, UK
miles@inf.ed.ac.uk

Mark Dredze

Human Language Technology Center of Excellence
Johns Hopkins University
Baltimore, MD 21211
mdredze@cs.jhu.edu

Abstract

Twitter is widely seen as being the go to place for breaking news. Recently however, competing Social Media have begun to carry news. Here we examine how Facebook, Google Plus and Twitter report on breaking news. We consider coverage (whether news events are reported) and latency (the time when they are reported). Using data drawn from three weeks in December 2013, we identify 29 major news events, ranging from celebrity deaths, plague outbreaks to sports events. We find that all media carry the same major events, but Twitter continues to be the preferred medium for breaking news, almost consistently leading Facebook or Google Plus. Facebook and Google Plus largely repost newswire stories and their main research value is that they conveniently package multiple sources of information together.

Introduction

Amongst researchers, Twitter is a very popular data source when finding breaking news, event detection and other interesting stories (see the survey (Atefeh and Khreich 2013)). In part this is because it is seen as having a real-time quality to it, and in part this is because the data is easily accessible via the freely available streaming API (which supplies a 1% random sample of posts). As valuable as this Twitter-centric body of research is, questions remain.

What are the coverage limitations of our results? One of the strengths of Twitter is that it enables citizens to report on news. But if these reports are missed from the sample, then they will go unnoticed (Morstatter et al. 2013). Are we even looking at the right Social Media? A recent survey of 5,173 adults suggested that 30% of people get their news from Facebook, while only 8% receive news from Twitter and 4% from Google Plus (Mitchell, Holcomb, and Page November 2013). There is clearly a mismatch between where academic researchers focus and where people go for news.

Twitter is not the only Social Media and the community needs to look beyond it to make our work on news and event detection more robust and relevant given where people actually consume Social Media news. While in the context of event detection there have been a few research efforts looking at other streams – such as Wikipedia or Flickr (Osborne et al. 2012; Chen and Roy 2009) – researcher have ignored

the elephant in the room: *Facebook*. With on average 727 million daily users (September 2013), it is the largest and most active Social Media.¹ Although Facebook is generally seen as being a medium for *private* communication between users, it is increasingly taking on a news reporting capacity. Users can *publicly* post updates that are indexed by major search engines and thus reach a large audience. Facebook itself has recognized that people prefer to read about high quality content, rather than just memes and cat photos.² With the introduction of the Graph Search API, researchers have access to Facebook posts. Similarly, while *Google Plus* has significantly fewer users, it provides an API for data access.

We go beyond most breaking news and event detection research and for the first time compare Facebook and Google Plus with Twitter. Additionally, since very little is known about Facebook and Google Plus, we analyze and compare data obtained from all three platforms. We address three research questions:

1. Do Social Media streams cover different events?
2. Where does news appear first?
3. What are data differences between the platforms?

With respect to major events, our results suggest that all Social Media streams largely cover the same events and that in general, they all lag behind traditional newswire and blog posts. Twitter is the most timely, followed by Facebook, with Google Plus in last place. Our results echo and generalize previous research, which considered the relation between newswire and Twitter, using data from 2011 (Petrovic et al. 2013). Facebook and Google Plus are perhaps best seen as extending the coverage of event detection when using the publicly available 1% Twitter streaming API. The richer environment and usage differences of Facebook and Google Plus point the way to new research opportunities.

Methodology and Datasets

For our comparison of event reporting, we take a two-stage approach: we consider when and whether our streams contain prespecified *major events* – events that any news service should carry – and whether they contain *long-tail events*

¹ <https://newsroom.fb.com/Key-Facts>

² <https://newsroom.fb.com/News/768/News-Feed-FYI-Helping-You-Find-More-News-to-Talk-About>

– events that might not be carried by traditional newswire and/or are not mentioned in our prespecified major event list.

Major Events

We use Wikipedia to identify 28 major events happening between 10th of December to 31st December 2013.³ These events (Table 4) are not tied to any particular medium and we would expect that any useful social stream would mention them. These events cover a broad range of categories, including those often associated with Social Media (namely natural disasters and celebrity deaths).

For each named event, we identified the first report in each stream, noting the UTC time of the post. We used Google (with the *site:* keyword) to manually search for events in each stream since Google indexes all publicly available posts, not just a sample. In some cases where we had difficulty finding early mentions of an event, we utilized Bing search as well as the search engines provided by Twitter, Facebook and Google Plus. For newswire, we used both Google News as well as Google and Bing. For cases when Social Media mentioned a newswire post that lacked a timestamp, or when the newswire story had a timestamp *after* the corresponding Tweet (presumably being an update time), we assumed the newswire post had the same time as the post (we found 5 such cases). Each event was investigated at least two times by two different people to ensure we found the earliest post. We note Google often did not index a specific post, instead indexing the timeline of the user, which often did not contain a relevant post. This made it challenging, particularly for Facebook posts, to find the relevant earliest post. In some cases, we manually reviewed dozens of posts from a single platform for a single event.

Long-tailed Events

We considered additional events not in our Wikipedia list by running a state-of-the-art event detection system over data crawled over the same interval (Petrovic, Osborne, and Lavrenko 2010). The system considers a post to be a newsworthy event if it appears new with respect to previously seen posts (in that stream) and has at least one closely related follow-up post appearing shortly afterwards in the same stream. The first part finds novel stories and the second part filters them, removing many false positives. This approach favors recall over precision, returning many spurious events.

Twitter data came from the standard streaming API. Facebook does not (to our knowledge) supply a random sample of posts, but it does provide a search API over public status updates, yielding 400 updates per query.⁴ To create a random sample status updates, we selected the top 1000 most frequent tokens from a month’s worth of Twitter data and continuously rotated between these tokens as queries to Facebook using a single crawler. This list included stop words in multiple languages, as well as a few Twitter-specific terms.

³http://en.wikipedia.org/wiki/December_2013

⁴While the API allows requests for more results, we found that larger requests often resulted in the API returning HTTP errors.

Stream	Property	Value
Twitter	Source	Twitter Streaming API
	Number of raw Tweets	97 Million
	Putative Events detected	3.3 Million
Facebook	Source	Graph Search
	Number of raw posts	7.8 Million
	Putative Events detected	228k

Table 1: Stream Statistics from December 10 to 31, 2013.

We treat the result as a randomly sampled stream.⁵ Note that we know nothing about which posts Facebook returns and cannot assume that the sample is in any way representative. For our analysis comparing the benefits of each platform we used a similar crawling strategy with the Google Plus API. However, given our named event reporting results we focused solely on Twitter and Facebook for automated event detection. See Table 1 for dataset summary statistics.

Do Social Media cover different events?

Previously work (Petrovic et al. 2013) established that Twitter and Newswire largely cover the same set of major events. Is this true for (public posts in) Facebook and Google Plus? Our major event results (Table 4) indicate that every event appears on every Social Media platform. Drilling-down, we noticed that the same content is often cross-posted. This echoes the way that news content is syndicated in traditional newswire; Social Media becomes just another publication medium.

Turning now to the long-tailed events, we consider whether Facebook contains events not mentioned in Twitter?⁶ Should those using Twitter data also consider Facebook data, or does Twitter contain every Facebook event?

For both Twitter and Facebook, we used the same event detection settings. Table 1 gives some statistics about the raw and filtered event streams. While Twitter provides considerably more data, both streams yield similar relative amounts of possible events (approximately 3%).

We then filtered the event streams using a classifier trained on approximately 100k manually labeled events (interesting vs. not interesting event) detected in Twitter over the Summer of 2011.⁷ Table 2 shows examples of content classification. This dramatically reduces the number of events, with a risk of false positives and a possible bias towards stories likely to appear in Twitter. The Twitter event stream reduces from 3.3 million possible events down to 6053 events, whilst the Facebook stream reduces from 228k down to 728 events. For each Facebook event, we identified the closest matching Twitter event. This attempts to pair corresponding stories. Finally, we sampled 100 event-pairs for inspection.

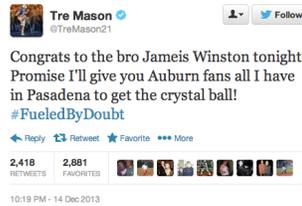
⁵Adaptive querying could obtain an unbiased sample from a static document set (Bar-Yossef and Gurevich 2008), whereas we have a dynamic set.

⁶We would expect that event detection using Google Plus to yield worse results, given that it has fewer users.

⁷Interesting, or newsworthy, events include bombings, takeovers, celebrity deaths, etc.



(a) Facebook



(b) Twitter



(c) Google Plus

Figure 1: Examples of first posts from each Social Media for the event “Jameis Winston wins Heisman trophy”. Twitter and Facebook were from Citizen Journalists (users), whereas Google Plus is a story from an official news outlet.

Event
Iran, world powers to hold nuclear talks http
#BREAKING Tsunami Alert for Japan as second earthquake hits coast of Japan at 7.9 Magnitude
Rupert Murdoch joins Twitter, immediately comes under fire http: by @m4tt on @twtwit
All of your relationships could seem complicated these days an... More for Virgo http:
@timeofy0urlife Are you interested in being your own boss? Take control now and start making 4k a month in 3 months
@shyB28 Want to be your own boss? I became mine 3 months ago and now I make 4k a month working fro home
You were born because you're going to be important to someone

Table 2: Example discovered events automatically classified as content-carrying (in **bold**) or spurious.

17 / 100 pairs discussed the same event. For example, both streams mentioned the Peter O’Toole death, the Target credit card data loss and the academic strike in Nigeria. This shows that the Facebook crawl does contain useful information. Looking in detail at 20 Facebook stories that did not have a corresponding matching Twitter story, we found that 13 had matching Twitter stories when considering the full Twitter firehose (searching Twitter.com), while two posts discussed stories not in Twitter: Taiwan grounding helicopters and an opinion piece about Russia. This shows that the Facebook crawl extends coverage of the Twitter crawl and to a lesser extent shows that there are stories in Facebook not reported in Twitter. The remaining stories were false positives.

In summary, to answer Question One: it appears that all Social Media sites cover the same major news events and probably largely cover the same long tail of events.

Where does news appear first?

Twitter has a reputation for being the first place to report on certain kinds of events, such as Earthquakes or Sports events. Out of the three Social Media, it consistently carries news before either Facebook or Google Plus. This is shown in Table 3, which measured the average (and stdev) relative lag: the time between first post overall and the stream’s first post. If Newswire carried a story at noon, Twitter at

Stream	Mean Latency	# Scoops (SM only)
Newswire	0.54 (9.26)	22
Twitter	2.36 (2.36)	9 (19)
FaceBook	9.89 (78.33)	2 (4)
Google Plus	14.01 (208.18)	0 (6)

Table 3: Mean (and stdev) reporting latency in hours (lower is better). Scoops are the number of global first reports for each stream (and Social Media only). Ties count for both; higher is better.

1pm, Facebook at 2pm and Google Plus at 3pm, the latencies would be 0, 1, 2 and 3 hours respectively. The Latencies show that Twitter reports news much faster than Facebook or Google Plus, but still lags newswire. This latter point updates previous work on 2011 data (Petrovic et al. 2013). Additionally, Twitter has nearly all the “scoops” (first reported post) of the Social Media, but much less than newswire. Facebook broke the news first for the Miss World story (via an account of a previous winner) and for a drug smuggling story – involving an Irish national – by an Irish Radio station. Aside from those two stories, Twitter led Facebook *in every other case but two*. Interestingly, Twitter did not always lead newswire for disasters. An event involving the ceiling falling in a theatre broke via newswire, which posted it soon after on Twitter.

In summary, to answer Question Two: Twitter is the best for breaking news, but still trails newswire.

What are data differences between platforms?

While all platforms provide similar news coverage, there are clear differences in the type of content. Table 4 compares properties of each platform. Even with a 1% cap, the Twitter streaming API provides more data. Facebook and Google Plus provide much richer posts, both in terms of length – Facebook posts are roughly 10 times as long as tweets, and Google Plus 5 times as long – and structured content, i.e. long chains of comments. While Twitter has a conversation feature, the 1% API is unlikely to supply all Tweets in a conversation, whereas Facebook and Google Plus posts include associated comments.

We observe several interesting differences in the posts. A huge number of Google Plus posts (44.2%) contain links,

Event	NewsWire	Twitter	Facebook	Google Plus
Uruguay legalizes cannabis	23:49 Dec 10	11:55 Dec 11	<i>01:21 Dec 11</i>	02:35 Dec 11
Pope person of year	11:50 Dec 11	<i>12:48 Dec 11</i>	12:55 Dec 11	16:03 Dec 11
Bubonic plague outbreak	17:24 Dec 10	19:30 Dec 11	18:00 Dec 12	<i>00:22 Dec 11</i>
Golden Globe nominations named	13:18 Dec 12	<i>13:22 Dec 12</i>	13:23 Dec 12	14:32 Dec 12
Jang Sung-taek executed	18:20 Dec 12	<i>21:35 Dec 12</i>	23:00 Dec 12	22:19 Dec 12
Google removes privacy feature	09:25 Dec 12	<i>09:48 Dec 12</i>	07:25 Dec 14	09:50 Dec 12
Car bomb in Mali	09:48 Dec 14	09:48 Dec 14	19:18 Dec 14	14:04 Dec 14
Peter O'Toole dies	17:17 Dec 15	<i>18:21 Dec 15</i>	21:20 Dec 15	01:18 Dec 16
Chinese on moon	13:38 Dec 14	13:38 Dec 14	16:39 Dec 14	17:42 Dec 14
Jameis Winston wins Heisman	02:01 Dec 15	03:19 Dec 15	03:59 Dec 15	<i>02:37 Dec 15</i>
Jane Fontaine dies	00:42 Dec 16	<i>01:05 Dec 16</i>	02:29 Dec 16	06:01 Dec 16
Google buys Boston Dynamics	06:27 Dec 14	08:40 Dec 14	05:04 Dec 15	<i>08:10 Dec 14</i>
Michelle Bachelet elected	22:09 Dec 15	03:56 Dec 16	05:38 Dec 16	<i>22:28 Dec 15</i>
RnR Hall of Fame inductions	04:25 Dec 17	<i>04:37 Dec 17</i>	07:43 Dec 17	13:07 Dec 17
Santiago wins Miss World	12:55 Dec 17	12:55 Dec 17	12:45 Dec 17	19:41 Dec 17
Two sentenced for drug smuggling	16:00 Dec 17	17:46 Dec 17	16:00 Dec 17	19:18 Dec 17
Angela Merkel reelected	14:03 Dec 17	09:24 Dec 17	13:44 Dec 17	07:21 Dec 18
Australia wins Ashes	05:45 Dec 17	06:14 Dec 17	06:24 Dec 17	<i>06:05 Dec 17</i>
Gonzalo Inzunza Inzunza killed	19:26 Dec 18	<i>20:26 Dec 18</i>	00:54 Dec 19	00:36 Dec 20
Pussy Riot given amnesty	22:03 Dec 9	13:36 Dec 9	18:12 Dec 11	12:43 Dec 18
Ronnie Biggs dies	07:43 Dec 18	07:04 Dec 18	12:45 Dec 18	10:33 Dec 18
Target credit cards compromised	22:07 Dec 18	22:07 Dec 18	13:01 Dec 19	22:39 Dec 18
Apollo Theatre ceiling collapses	20:30 Dec 19	<i>20:31 Dec 19</i>	21:08 Dec 19	09:40 Dec 20
Canada overturns prostitution laws	11:00 Dec 19	<i>17:20 Dec 19</i>	09:15 Dec 20	07:38 Dec 20
Alan Turing pardoned	21:19 Dec 23	21:19 Dec 23	22:36 Dec 23	23:24 Dec 23
Shinzo Abe visits Yasukuni shrine	02:11 Dec 26	01:52 Dec 26	05:15 Dec 26	03:22 Dec 26
Joaquin Guzman heart attack	08:15 Dec 26	<i>11:22 Dec 26</i>	00:10 Dec 27	17:41 Dec 28
MS King Seaways catches fire	23:23 Dec 28	22:58 Dec 28	23:43 Dec 28	06:27 Dec 29

Property	Twitter	Facebook	Google Plus
Obtaining Data			
Crawling Method	Streaming	Search	Search
Messages per day	4.2m	2.3m	180k
Messages			
Mean/Med./STD length (chars)	68/61/40	721/206/2105	433/206/950
Mean/Med./STD length (toks)	10/8/7	125/38/357	55/23/123
Contains link	16.0%	22.7%	95.4% (44.2% excluding albums)
Contains hashtag	14.2%	0.5%	0.04%
Contains username	55.8%	3.1%	5.0%
Contains image	10.4%	21.3%	58.2%
Resharing post	27.4%	68.4%	29.2%
Geolocated	2.9%	1.0%	0.6%
Platform			
Real Name Only	No	Yes	Yes
Feedback	favorites	Likes	+1
Discussion	Retweets ^a	Comments	Comments
Posting Method	iPhone (26.9%) Android (19.8%) Web (18.3%)	Android (15.3%) iPhone (14.4%) Mobile (12%)	NA

^aConversations are not available via the streaming API.

Table 4: a) UTC times of first posts. Bold: first post, italics: first Social Media. b) Properties of different Social Media (based on two days worth of crawled messages).

increasing to 95.4% when including links to Google albums. Google Plus is primarily used to share external content, as opposed to Twitter and Facebook. Hashtags are almost exclusively the domain of Twitter (Facebook (0.5%) and Google Plus (0.04%)).⁸ Twitter posts frequently contain references to other users (55.8%), likely a side effect of the way conversations are implemented. Over a quarter of Twitter and Google Plus posts are retweets/shares, while 68.4% of Facebook posts are shares.⁹ Twitter has the highest geolocation rate (2.9%), three times that of Facebook (1.0%). Others have greatly expanded Twitter geolocation using the location field in user profile (Dredze et al. 2013), whereas Facebook and Google Plus do not include this information.

Twitter does not have a real name requirement for users, whereas Facebook and Google Plus do. This may have an impact on spam, trust and author attribution. Both Facebook and Google Plus have a concept of "like", which opens an interesting research direction, looking at the relationship between public declarations and (for example) rumor detection, impact and message propagation.

In summary, to answer question three: each platform provides different features and is used differently, suggesting future work on how users perceive and use each platform.

Conclusion

We presented the first results for breaking news in Facebook and Google Plus. Our results show all public posts in Social

⁸When viewed in a web browser, many Google Plus posts contain hashtags, but these are not available using the API. These may be automatically inferred after the post is created.

⁹Computing this number is difficult since Facebook shares do not show up as separate posts. Instead, each post contains a number indicating how many times it was shared at the time of crawl. Summing this number yields a sharing of 68.4%, and 1.1% of posts are shared at least once.

Media streams carry similar events to each other. Yet Twitter dominates other Social Media in providing timely news. Still, Twitter lags news wire, which remains the best source for breaking news. Despite Twitter's dominance, we demonstrate that these other platforms offer interesting possibilities for many interesting research directions.¹⁰

Acknowledgements MO acknowledges support from EP-SRC/DSTL grant EP/L010690/1.

References

- Atefeh, F., and Khreich, W. 2013. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence n/a-n/a*.
- Bar-Yossef, Z., and Gurevich, M. 2008. Random Sampling from a Search Engine's Index. *Journal of the ACM (JACM) 55(5):24*.
- Chen, L., and Roy, A. 2009. Event Detection from Flickr Data through Wavelet-based Spatial Analysis. In *CIKM*.
- Dredze, M.; Paul, M. J.; Bergsma, S.; and Tran, H. 2013. Carmen: A twitter geolocation system with applications to public health. In *HIAI*.
- Mitchell, A.; Holcomb, J.; and Page, D. November 2013. News use across Social Media Platforms. Technical report, Pew Research Center.
- Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*, 400–408. AAAI.
- Osborne, M.; Petrović, S.; McCreadie, R.; Macdonald, C.; and Ounis, I. 2012. Bieber no more: First Story Detection using Twitter and Wikipedia. In *Proceedings of the SIGIR workshop on Time-Aware Information Access*.
- Petrovic, S.; Osborne, M.; McCreadie, R.; Macdonald, C.; Ounis, I.; and Shrimpton, L. 2013. Can Twitter replace NewsWire for Breaking News? In *ICWSM*.
- Petrovic, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story detection with application to Twitter. In *NAACL '10*.

¹⁰More details, such as data from Table 4, can be found at: <http://goo.gl/6ZCec3>.