Can Twitter replace Newswire for breaking news?

Saša Petrović[*] **Miles Osborne**[*] **Richard McCreadie**[**] **Craig Macdonald**[**] **Iadh Ounis**[**] **Luke Shrimpton**[*]

School of Informatics[*] School of Computer Science[**]
University of Edinburgh University of Glasgow
contact: miles@inf.ed.ac.uk

Abstract

Twitter is often considered to be a useful source of real-time news, potentially replacing newswire for this purpose. But is this true? In this paper, we examine the extent to which news reporting in newswire and Twitter overlap and whether Twitter often reports news faster than traditional newswire providers. In particular, we analyse 77 days worth of tweet and newswire articles with respect to both manually identified major news events and larger volumes of automatically identified news events. Our results indicate that Twitter reports the same events as newswire providers, in addition to a long tail of minor events ignored by mainstream media. However, contrary to popular belief, neither stream leads the other when dealing with major news events, indicating that the value that Twitter can bring in a news setting comes predominantly from increased event coverage, not timeliness of reporting.

Introduction

Twitter today is becoming a *de facto* standard domain for event detection (Becker, Naaman, and Gravano 2011; Li, Sun, and Datta 2012; Li et al. 2012; Weng et al. 2011). Indeed, due to its real-time nature, Twitter has can be used as a sensor to gather up-to-date information about the state of the world. For example, Sakaki *et al.* (Sakaki, Okazaki, and Matsuo 2010) use Twitter for early detection of earthquakes in hope of sending word about them *before* they even hit, while Demirbas *et al.* (Demirbas et al. 2010) use Twitter users as sensors to get the current weather conditions.

One of Twitter's main strengths that is often cited in the literature is its real-time nature, i.e. that users post about events as they are happening. In fact, it is often suggested that Twitter breaks news before newswire, with a handful of examples cited to support this claim (Kwak et al. 2010; Sakaki, Okazaki, and Matsuo 2010). However, there has been relatively little work to substantiate these claims. Instead, most prior work that does exist in this area focuses on comparing newswire and tweet streams in terms of topical similarity, ignoring the time aspect. For instance, (Subašić and Berendt 2011) compared tweets and blogs to articles from Reuters, Associated Press (AP), and other professional news outlets in terms of the similarity of the underlying language models. They found that tweets were textually very similar to the headlines of related news articles, but also that those same tweets were dissimilar to the full article content. The only work that we are aware of which considers the time aspect of news reporting in the two streams is by (Kwak et al. 2010), where the authors compare CNN and Twitter. However, they provide no quantitative analysis or discussion, stating only that CNN mostly leads Twitter and that some events such as sports matches and accidents break on Twitter sooner.

In this paper, we perform a detailed analysis of major and minor events reported during a 2 month period in newswire and tweet streams. Our aim is to determine whether Twitter is actually now able to replace traditional newswire providers in terms of coverage and timeliness of news reporting. To this end, we analyse two sets of events. First, a small set of manually identified events relating to major news stories from the time. Second, a larger set of potential events generated using an automatic event detection algorithm. For both sets of events, we investigate three research questions, namely:

- Do newswire providers and Twitter cover different events? (RQ1)
- Does Twitter report news faster than traditional newswire providers? (RQ2)
- What types of events does Twitter report first? (RQ3)

When considering major news events, our results suggest that both traditional newswire providers and Twitter report on major events, but that Twitter has better coverage of smaller or local events that are typically ignored by the newswire sources we examined. Furthermore, in contrast to popular belief, our results show that neither stream consistently reports on breaking news first. Indeed, traditional newswire sources often report events before Twitter.

Methodology and Datasets

To facilitate our comparison of events reported by newswire providers and Twitter, we develop a new dataset containing newswire and tweet streams from a period of time, in addition to events extracted from that time. In particular, we use newswire and tweet streams crawled from the period of June 30^{th} 2011 to September 15^{th} 2011 (77 days). Table 1 reports the statistics of these two streams. Notably this period was

Stream	Property	Value
Newswire	Sources	BBC, CNN, Google News, New York Times Guardian, Reuters, The Register, and Wired
	Time-Range	30/06/2011 → 15/08/2011
	# Articles	47,000
	# Clustered Events	27,000
Twitter	Source	Twitter Streaming API
	Time-Range	30/06/2011 → 15/08/2011
	# Tweets	51,000,000
	# Clustered Events	25,000,000 (800,000 w/o singletons)

Table 1: Stream Statistics.

chosen since it has been used in prior work (Petrović, Osborne, and Lavrenko 2012) and is known to contain a wide variety of interesting news events.

Next, we generated two sets of events, one manually and one using automatic event detection software. In particular, our first event set is comprised of 27 manually identified events, each relating to a major news story from the period of our dataset. A listing of these events can be found later in Table 4. These events cover a range of categories, from celebrity news to accidents, and from natural disasters to science discoveries. Note that these events are unbiased – they are not restricted to some geographical locality, nor are they solely planned or somehow connected to the behaviour of some system[1].

However, this is a precision-oriented approach that lacks coverage of many (potentially interesting) events. Hence, we generate a second event set automatically using a state-of-the-art event detection/clustering system (Petrovic, Osborne, and Lavrenko 2010). Figure 1 illustrates how we generate events automatically from our newswire/tweet streams. From Figure 1, we first cluster both tweets and newswire into two larger sets of potential events, i.e. events reported in Twitter and events reported by newswire providers. Each cluster within a set corresponds with a potential new event and may contain associated follow-up posts. Table 1 reports the number of event clusters extracted from each stream. From Table 1 we see that there are many event clusters generated from each stream. Having extracted potential events from each stream, we try to find events that are reported in both streams. We do this by finding, for each newswire (Twitter) cluster, its nearest neighbour in terms of cosine similarity in the stream of Twitter (newswire) clusters. We refer to the cosine similarity between the each cluster and its nearest neighbour as the *pair similarity*. In both streams, clusters are represented by the sum of all document vectors in the cluster. Note that, because most of the 25 million clusters in Twitter consist only of one tweet, we discard those clusters as spurious, and in the rest of our experiments we work only with the 800,000 non-singleton Twitter clusters.

Event Coverage in Newswire and Twitter

In this section, we investigate the first of our three research questions, i.e. Do newswire providers and Twitter cover different events? To answer this question, we take a three-stage approach. First, we examine the proportion of the 27 major events identified previously that both newswire and Twitter streams reported on. Second, we examine the overlap between the automatically identified events from newswire to those in Twitter. Third, we examine the reverse case, i.e. events reported in Twitter but not in newswire.

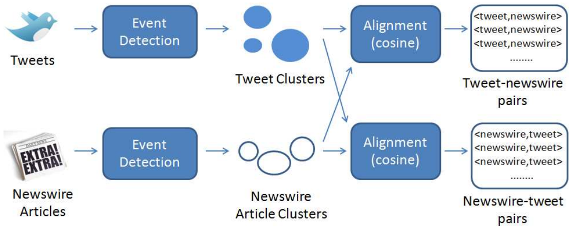

Figure 1: Methodology for the automatic generation of tweet (cluster) and newswire article (cluster) pairs, representing potential events.

As a sanity check, we begin by examining the 27 major events we identified in the previous section. For each event, we manually searched starting from the time when the event actually happened with the aim of finding newswire articles and tweets about each article, respectively. From this analysis, we found that all 27 events were reported by newswire providers and Twitter.

Next, to investigate the overlap of events covered by each stream in more detail, we use both the news-tweet and tweet-news pairs (see Table 1). In particular, we focus on those pairs with a low similarity, i.e. those newswire events where a good match was not found in Twitter. Using a similarity threshold, we selected the 3,000 news-tweet pairs with the lowest similarity. We then randomly sampled 100 of these and labelled them according to whether those pairs actually refer to the same event or not. We found that 79 of those pairs do refer to the same event, whereas 21 did not. Because the alignment algorithm is not perfect, we further inspect the 21 news events manually. In particular, we manually search our dataset of 51 million tweets using keyword search, looking for mentions of those 21 events. Using this manual keyword search, we find that 16 of those events were mentioned in Twitter but the alignment algorithm did not find them, whereas five were not mentioned at all in our Twitter data. The five events not found in our data are shown in Table 2. We can see that, while the first two are actual events, the other three do not correspond to real events, and are more analysis-type articles. Overall, this shows that even the public 1% stream of Twitter contains around 95% of all the events reported in the newswire, i.e. Twitter has indeed a very good coverage of newswire events.

We now consider whether events reported in Twitter also get reported in the newswire, i.e. we examine our tweet-news pairs (rather than news-tweet pairs). It is clear from the orders of magnitude more posts submitted to Twitter (see Table 1) that it contains much more information than newswire. However, when it comes to reporting news, it is not clear if Twitter actually contains information that newswire does not. As with the previous experiment, we assume that tweet-news pairs with a high similarity represent events posted in both streams and focus on those with low similarity. However, Twitter generates a very large number of potential

[1] The dataset that is used in this paper can be obtained from *http://demeter.inf.ed.ac.uk/cross/*

Iranian actor Pegah Ahangarani arrested in Teheran
Court upholds decision to impose control order on terror suspect in London
Support of Assad government shows signs of weakening
French socialist project ’sharing and caring’ in bid to beat Nicolas Sarkozy
Dick Cheney autobiography heaps praise on Tony Blair

Table 2: News events not found in our Twitter data.

events, of which many will be noise, making it impractical to label all pairs with a low similarity. Instead, we randomly sample 1,000 events and label them as corresponding to real events or not, which will give us an estimate of how many real events exist in Twitter that are not covered in the newswire. We also manually inspect the newswire data using keyword search and realign cases where the correct alignment was not found by the algorithm.

Table 3 shows examples of the events that were reported on Twitter but not in the newswire stream. Out of 1,000 potential Twitter events, 54 were actual events (determined by manual labelling), while the rest were noise. Out of those 54 events, we found that 28 events (52%) were not reported in the newswire. 15 of these events were sports-related, while the rest were a mix of all other event types. This is in line with the findings of (Kwak et al. 2010), who noted that Twitter leads CNN mostly in sports events, and suggests that Twitter is a very good source of up-to-date sports news, probably because a lot of sports fans tweet about the games as they unfold. The fifth example in Table 3 shows another strength of Twitter. During the London riots, there were a lot of tweets about minor acts of vandalism that did not make it into the mainstream news, but Twitter served as a medium that carried all of these hyper-local events. The last example in Table 3 also shows how the law enforcement used Twitter to dispute rumours during the riots. This information would not make it into newswire as it only had value for a limited time period, making Twitter an ideal medium to carry it.

To answer our first research question, Twitter appears to cover nearly all newswire events, but newswire only covers a subset of the events reported in Twitter. Most of the Twitter-only events are sports-related, have value only for a short time or to a very restricted audience, and would thus lose value by the time they are reported in the newswire.

Timeliness of News Reporting in Newswire and Twitter

In this section, we examine our second research question, i.e. does Twitter report news faster than traditional newswire providers? In particular, we first investigate the timeliness of event reporting for each of the 27 major events we manually identified. We then analyse our automatically identified events to estimate upper and lower bound for the proportion of events reported first in Twitter.

We begin by manually inspecting the set of 27 (high-profile) reference events in order to establish the time difference of the first mention of these events in Twitter and in newswire. For Twitter, we manually search our tweets, starting from the time when the event actually happened. For newswire, we conduct a thorough search of the Web, looking for the first newswire article corresponding with each event. In doing so, we consider a much larger volume of newswire than our original eight sources.

Event	Newswire	Twitter	Lead
Amy Winehouse dies	**07-23 16:10**	07-23 16:11	-0:01
Atlantis shuttle lands	07-21 09:59	**07-21 09:56**	+0:03
Betty Ford dies	**07-09 00:00**	07-09 00:57	-0:57
Richard Bowes killed in riots in England	**08-11 23:18**	08-11 23:31	-0:14
Flight 4896 crash	**07-13 11:37**	07-13 11:46	-0:09
S&P downgrade US credit rating	**08-06 00:11**	08-06 00:18	-0:07
US increases debt ceiling	08-01 23:06	08-01 23:06	0:00
Terrorist attack in Delhi	09-01 05:12	**09-07 04:53**	+0:19
Earthquake in Virginia	08-23 18:24	**08-23 17:53**	+0:31
First victim of London riots dies	08-09 11:46	**08-09 11:45**	+0:01
War criminal Goran Hadzic arrested	**07-20 07:56**	07-21 05:42	-21:46
India and Bangladesh sign a border pact	**09-06 07:15**	09-06 14:24	-7:09
Plane with Russian hockey team Lokomotiv crashes	**09-07 12:51**	09-07 12:59	-0:08
Explosion in French nuclear plant in Marcoule	09-12 11:42	09-12 11:42	0:00
NASA announces there might be water on Mars	08-04 18:08	08-04 18:08	0:00
Google announces plans to buy Motorola Mobility	08-15 11:43	**08-15 11:38**	+0:05
Car bomb explodes in Oslo, Norway	07-22 13:57	**07-22 13:38**	+0:19
Gunman opens fire in youth camp in Norway	**07-22 16:13**	07-22 16:14	-0:01
First artificial organ transplant	**07-07 16:03**	07-07 16:25	-0:22
Petrol pipeline explodes in Kenya	**09-12 04:34**	09-12 08:17	-3:43
Famine declared in Somalia	07-20 07:21	07-20 07:21	0:00
South Sudan becomes independent country	**07-08 21:03**	07-08 21:05	-0:02
South Sudan becomes UN member state	**07-14 14:23**	07-14 14:31	-0:08
Three men die in riots in England	08-10 06:33	**08-10 05:45**	+0:48
Riots break out in Tottenham, England	08-06 21:13	**08-06 20:08**	+1:05
Rebels capture International Tripoli Airport	**08-21 08:00**	08-21 23:08	-15:08
Ferry sinks in Zanzibar	**09-10 04:21**	09-10 06:56	-2:35

Table 4: Times (in UTC) of events, first newswire stories, first tweets and lead (+ when Twitter leads).

Table 4 reports the delay between each of the 27 events being reported in Twitter and a corresponding newswire article that we found. The columns show the time of the first mention in newswire , the time of the first mention in Twitter and the how much Twitter leads newswire. Entries marked in bold occur first. From Table 4, we observe that Twitter unambiguously leads newswire eight times, newswire leads Twitter 15 times and the news breaks at the same time for four events. This indicates that, contrary to popular belief, neither stream dominates the other in terms of high-profile breaking news.

It is possible that Twitter might lead for less well-known events, given the large volume of content posted. To examine this, we use the larger set of 27,000 news-tweet pairs (representing newswire events and their closest Twitter event). We first sort the news-tweet pairs according to their pair similarity and keep the 50% with the highest similarity. This way we remove pairs where the news and tweet clusters are not likely to refer to the same event. We then remove all pairs where the event was reported in newswire before Twitter, i.e. where the timestamp for the newswire article is older than that for the tweet. This leaves approximately 5,500 events that were reported in Twitter before newswire, or approximately 20% of cases. The median lead time of Twitter for these events was 53 minutes.

This provides an upper-bound estimation of the number of events for which Twitter leads. However, newswire sources other than the eight in our dataset (see Table 1) might have posted before Twitter and there is also no guarantee that our news-tweet alignments are correct. Hence, to generate a refined estimation, we use a series of filtering steps to remove news-tweet pairs that should not be considered as cases where Twitter broke the news about the event first. In particular, we remove tweets with a link (indicating that the news has already been reported), remove pairs where the

RIP Rick Rypien. Sad to see another death in the NHL. Too many tragedies in the world lately...
UFC on Versus 5 results: Jacob Volkmann def. Danny Castillo via unanimous decision (29-28, 29-28, 29-28)
RT @NASA: NASA is ready to move forward with Space Launch System, a new capability for human exploration beyond Earth
Car reg NP05 LPU looting PC World Charlton. Retweet and shame.
RT @DerbysPolice: To reiterate rumours circulating there is disorder or looting in Derby city are untrue. Please RT. #derby #police

Table 3: Examples of events reported on Twitter, but not in our newswire data.

Magnitude 5.4 earthquake hits western Japan
Rapper Lil Wayne ends up in hospital after a skateboarding accident
Malaysian police use tear gas on protesters
Baidu senior VP resigns
Sherwood Schwartz dies
Thor Hushovd wins stage 13 of Tour de France
Japan wins FIFA Women's World Cup
Michele Bachmann wins Iowa straw poll

Table 5: Examples of events where Twitter leads newswire.

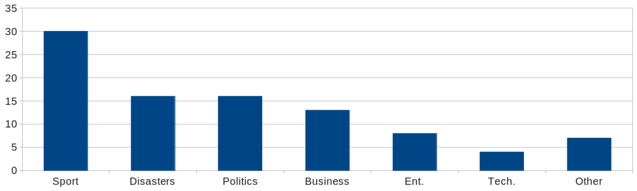

Figure 2: Events categories where Twitter leads.

tweet mentioned a newswire source, remove retweets, remove tweets copied from newswire, remove duplicate events and finally we manually realigned pairs that were incorrect, (miss-aligned pairs are dropped). This leaves us with only 97 news-tweet pairs, that we are confident represent events that broke on Twitter before newswire, or less than 1% of the 27,000 events we started with.

To answer our second research question, for major events neither stream consistently leads the other in terms of reporting time, i.e. both sources can report news first. Meanwhile, for the automatically detected events, there are very few where Twitter leads newswire when both sources report a story. Examples of these events are shown in Table 5.

Event Types Reported First on Twitter

Finally, we examine our third research question, i.e. What types of event does Twitter report first? To answer this question, we manually categorise the 97 automatically detected events from the previous section where Twitter reported before newswire into seven broad categories, namely: politics, sports, disasters & accidents,[2] business & economy, entertainment, technology, and other. These categories are motivated by TDT's broad topic types (Allan 2002) (also known as *rules of interpretation*).

Figure 2 shows the number of events that belong to each category. From Figure 2, we can see that many of the events when Twitter leads are sports events (one third) and disaster-related events. However, it is interesting to note that there are also cases when Twitter leads for politics and business-related events, where we would expect the newswire to carry first. Hence, to answer our third research question, Twitter tends to lead for sporting-related events and unpredictable high-impact natural disasters.

Conclusion

In this paper, we asked whether Twitter and newswire providers report on the same events and whether either source prominently leads the other. Our study has suggested that Twitter covers most of the events that are reported by newswire providers and that many events reported in Twitter are not mentioned in newswire. However, for the events that both streams report on, there is no evidence that one source leads the other in terms of breaking news. In general, while Twitter can break news before newswire in limited cases, for major events there is little evidence that it can replace newswire providers. On the other hand, the greater coverage of hyper-local news observed within Twitter supports the idea that it can be used for localised use-cases such as community policing or local search.

Acknowledgements

The authors acknowledge financial support for the CROSS Project from EPSRC grant EP/J020664/1.

References

Allan, J. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, 2002.

Becker, H.; Naaman, M.; and Gravano, L. Beyond trending topics: Real-world event identification on Twitter. In *Proc. of WSM*, 2011.

Demirbas, M.; Bayir, M. A.; Akcora, C. G.; Yilmaz, Y. S.; and Ferhatosmanoglu, H. Crowd-sourced sensing and collaboration using Twitter. In *Proc. of WoWMoM*, 2010.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. What is Twitter, a social network or a news media? In *Proc. of WWW*, 2010.

Li, R.; Lei, K. H.; Khadiwala, R.; and Chang, K. C.-C. TEDAS: A Twitter-based event detection and analysis system. In *Proc. of ICDE*, 2012.

Li, C.; Sun, A.; and Datta, A. Twevent: Segment-based event detection from tweets. In *Proc. of CIKM*, 2012.

Petrovic, S.; Osborne, M.; and Lavrenko, V. Streaming first story detection with application to Twitter. In *Proc. of NAACL*, 2010.

Petrović, S.; Osborne, M.; and Lavrenko, V. Using paraphrases for improving first story detection in news and Twitter. In *Proc. of NAACL*, 2012.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. of WWW*, 2010.

Subašić, I., and Berendt, B. Peddling or creating? Investigating the role of Twitter in news reporting. *Inf. Retrieval*, 2011.

Weng, J.; Yao, Y.; Leonardi, E.; and Lee, F. Event detection in Twitter. In *Proc. of WSM*, 2011.

[2] Also contains events like terrorist attacks or shootings.