The News Cycle's Influence on Social Media Activity

Andrew Yates
Information Retrieval Lab
Department of Computer Science
Georgetown University
andrew@ir.cs.georgetown.edu

Jonah Joselow
Information Retrieval Lab
Department of Computer Science
Georgetown University
jonah@ir.cs.georgetown.edu

Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
nazli@ir.cs.georgetown.edu

Abstract

While much work has studied the problem of identifying real-world trends based on social media, none has attempted to explicitly model the news cycle's influence on this social media activity. In this work we attempt to model the news cycle's influence on Twitter activity in the context of "news-centric events." We present a model for estimating the number of tweets posted in response to a news event and propose a method for creating an appropriate ground truth. We find that, although our method is sensitive to variations in the amount of training data, we are able to predict future Twitter activity with reasonable accuracy.

Introduction

Much work has studied the problem of using social media (e.g., tweets) and Web search engine logs to predict real-world outcomes in applications such as consumer behavior (Goel et al. 2010), election results (Tumasjan et al. 2010), public opinion (O'Connor et al. 2010), movie box office revenues (Asur and Huberman 2010; Mishne and Glance 2005; Oghina et al. 2012), and the prevalence of influenza-like illnesses (ILI) in a geographic area (Achrekar et al. 2011; Paul and Dredze 2011; Copeland et al. 2013; Signorini, Segre, and Polgreen 2011; Polgreen et al. 2008; Dugas et al. 2013). Previous work did not attempt to explicitly model the news cycle's influence on the user activity being measured (e.g., Web searches or tweets posted), yet research has suggested that the news cycle can cause such estimates to be inaccurate by causing a "celebrity effect" where users search or post about topics currently being covered by the news. (Cooper et al. 2005; Polgreen et al. 2008)

We propose a methodology for estimating the number of tweets posted in response to the news cycle, and show that a relationship exists between the number of tweets written and the number of news articles published about an event. Our contributions are: *(1)* a method for estimating the number of news-related tweets posted about an event, *(2)* a method for creating an appropriate ground truth consisting of news-related tweets about an event, and *(3)* an evaluation of how well our method can estimate the number of tweets posted in response to the news cycle (i.e., news-related tweets).

Methodology

Our goal is to predict the number of news-related tweets that will be posted about an event on a given day (i.e., the number of tweets that are posted in response to the news cycle). We create a ground truth consisting predominantly of news-related tweets by carefully choosing events with little Twitter activity before the events are covered by the news. Using features derived from the news cycle, we train a regression model to estimate how many tweets will be posted on a day d based on the news cycle leading up to day d.

Ground Truth

We create our ground truth by identifying news stories and tweets about "news-centric events," which are events characterized by a stable amount of low Twitter activity before news coverage begins and after news coverage ends. That is, the vast majority of tweets about news-centric events must be caused by the news coverage itself. Furthermore, the number of people these events affect directly should be small so that only an insignificant number of people are in a position to post tweets that are not a response to news coverage. News-centric events are by definition essentially absent from Twitter until the events take place. Natural disasters, murders, and unexpected disease outbreaks with a low prevalence (e.g., meningitis) are examples of such events; many people may tweet in response to such an event, but only a small number of people experienced the actual event.

We utilized Pearson Education's Information Please (Pearson Education 2015), an annual almanac, to identify news-centric events and retrieved related news articles from a variety of US news agencies' websites[1]. Using historical tweets collected from Twitter's 1% Streaming API, we identified tweets related to the events by manually choosing high-precision keywords that characterized the events, such as the names of the bombers in the case of the Boston Marathon Bombing event. We excluded any event that did not have significant Twitter activity, was not associated with high-precision keywords that could be used to identify tweets related to the event, or was an event that could be anticipated far in advance (e.g., seasonal influenza outbreaks).

[1] ABC News, CBS News, CNN, Fox News, The Huffington Post, NBC News, and NPR

Event Name	Days	News Data				Twitter Data				Spearman Corr.
		Articles	Mean	Max	σ	Tweets	Mean	Max	σ	
BP Oil Spill	44	879	20.0	46	12.6	42695	970.3	3911	863.5	0.60
Yosemite Hantavirus	79	63	0.8	7	1.6	518	6.6	87	15.8	0.47
2012 Meningitis Outbreak	182	314	1.7	10	2.4	2874	15.8	120	24.3	0.46
Hurricane Sandy	284	1433	5.1	126	12.8	46505	163.8	11622	898.0	0.63
Boston Marathon Bombing	89	1513	17.0	127	28.5	13875	155.9	3341	489.3	0.95
2010 Haiti Earthquake	69	978	14.2	88	19.6	26896	389.8	6284	962.1	0.86
2013 Midwest Tornadoes	42	575	13.7	116	23.0	35828	853.1	10364	1690.56	0.79
Sandy Hook Shooting	242	2352	9.7	154	18.3	67248	277.9	8276	914.4	0.50

Table 1: Summary statistics for the events included in our ground truth. Mean and max numbers are the number of articles or tweets per day. Spearman's ρ is calculated between the number of articles posted and the number of tweets posted on each day.

Event Name	NRMSE				
	$n = 10\%$	$n = 20\%$	$n = 30\%$	$n = 40\%$	$n = 50\%$
BP Oil Spill	0.314	0.332	0.185	0.197	0.232
Yosemite Hantavirus	0.206	0.213	0.182	0.040	0.036
2012 Meningitis Outbreak	0.226	0.185	0.119	0.094	0.095
Hurricane Sandy	0.083	0.061	0.017	0.012	0.008
Boston Marathon Bombing	0.161	0.171	0.183	0.157	0.045
2010 Haiti Earthquake	0.173	0.101	0.061	0.042	0.040
2013 Midwest Tornadoes	0.221	0.203	0.202	0.233	0.138
Sandy Hook Shooting	0.120	0.128	0.132	0.127	0.073

Table 2: The NRMSE for each event trained using up to $n = 50\%$ of the event. For most events accuracy improves significantly as more data is added, with an average NRMSE of approximately 0.08 at $n = 50\%$.

The list of events chosen and their keywords are available on the authors' website.[2]

Our high-precision keyword approach differs from recall-oriented prior work (Glasgow, Fink, and Boyd-Graber 2014). Rather than attempting to identify all tweets about an event, we attempt to identify high-precision keywords that characterize the event by identifying keywords that are *(1)* unlikely to refer to any other simultaneous event and *(2)* likely to remain relevant for the majority of the event. While our approach does not include all tweets related to each event, we note that relative differences in different days' tweet counts are more important to our model than absolute counts.

Table 1 shows the events included in our ground truth. The events vary greatly in terms of both duration and magnitude, from 63 news articles and 518 tweets about the 2012 Yosemite Hantavirus outbreak to 2,352 news articles and over 67,000 tweets about the Sandy Hook Shooting. Event durations were determined by considering the activity surrounding each event; Hurricane Sandy, for example, was mentioned in tweets and in news articles some time after the hurricane had passed (e.g., in the contexts of recovery progress and storm protection plans). While all events exhibit a positive rank correlation between the number of news articles published and the number of tweets authored, the magnitude of the correlations vary greatly across events, ranging from $\rho = 0.46$ to $\rho = 0.95$.

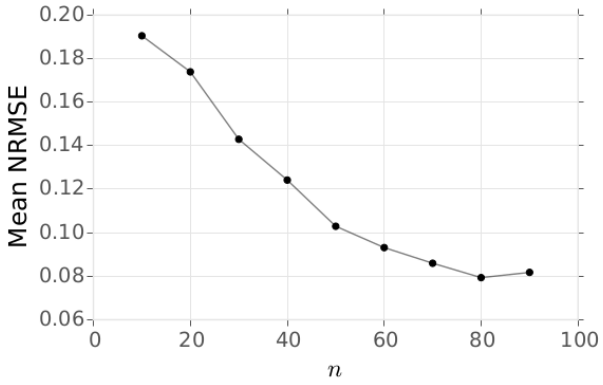

Figure 1: The mean normalized root-mean-square error across all events. The mean NRMSE continuously decreases as the model is trained on more data, but performs reasonably well when only the first 20% or 30% of an event is used. Lower NRMSEs are correlated with more news articles and tweets about an event.

[2] *http://ir.cs.georgetown.edu/data/news_icwsm16*

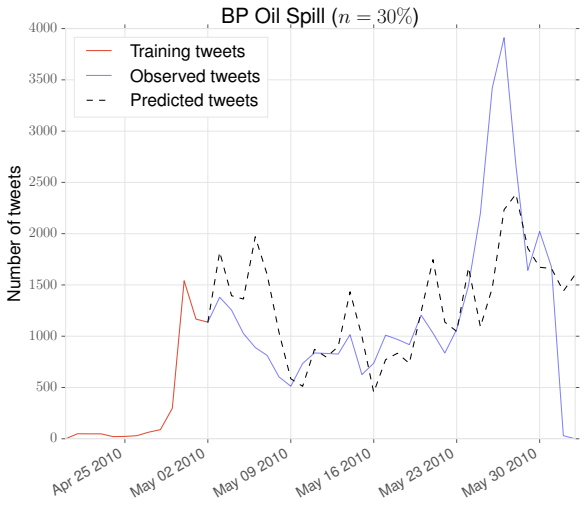

Figure 2: Prediction performance on the BP Oil Spill when the first 30% of the event is used as training data. Predictions follow the general trend of the observed tweets, but increase during a downward trend in early May and underestimate the Twitter activity in late May.

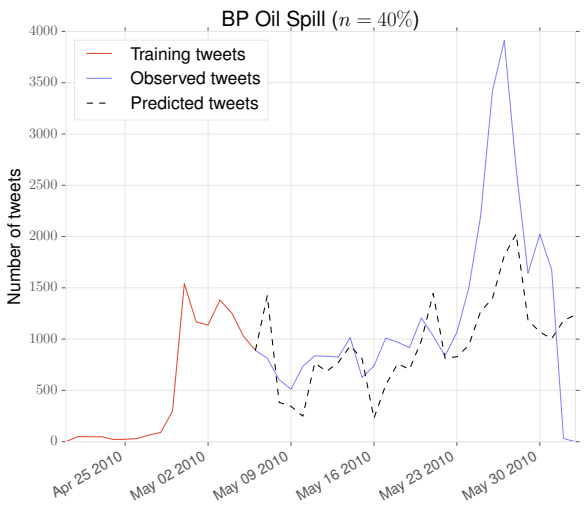

Figure 3: Prediction performance on the BP Oil Spill when the first 40% of the event is used as training data. Predictions are still substantially below the actual number of tweets posted in late May.

Model

We trained a regression model to predict the number of tweets posted on a day d using features calculated over rolling periods of time at least one day prior to d. We settled on using an ϵ-SVR (Support Vector Regression) as implemented by scikit-learn[3] with the default parameter values of $C = 1.0$ and $\epsilon = 0.1$.

We derive features from news articles about the target event. While we experimented with also including tweet features (e.g., the number of tweets posted per day, the sentiment and sentiment strength of tweets, etc.), such features did not significantly improve our model's performance. More importantly, using such features conflicts with the future goal of estimating the number of tweets posted in response to news about non-news-centric events (e.g., influenza-like illnesses), because in the case of such events we cannot know which tweets are news-related. Our ground truth events are specifically chosen to be news-centric events where we can safely assume that all tweets are news-related.

For each day d we incorporated the following features derived from news articles: *(1)* the total number of news articles published about the event on day d; *(2)* the maximum, minimum, mean, median, mode, and standard deviation of the number of news articles published about the event over a n-day rolling window ending on day d; *(3)* the overall sentiment polarity of news articles published about the event on day d (as determined by SentiWordNet (Baccianella, Esuli, and Sebastiani 2010)); and *(4)* the maximum, minimum, mean, median, mode, and standard deviation of the overall sentiment polarity of news articles published about the event over a n-day rolling window ending on day d.

[3] *http://scikit-learn.org*

These features are based on the hypotheses that *(1)* there is a relationship between the amount of news coverage of a news-centric event and the number of tweets written about the event (as is illustrated by the correlations in Table 1), and *(2)* the sentiment in news articles influences how much people tweet about the event. We experimented with using the different feature types in isolation, and found that including the sentiment features lowered our mean error by approximately 3%. While it is unlikely that all of the summary statistics generated over rolling windows (maximum, minimum, mean, etc.) are equally important, the regularization performed by the ϵ-SVR makes it unnecessary to manually select a subset of the features to utilize. We experimented empirically with different rolling window lengths and settled on 5-day rolling windows (i.e., $n = 5$).

Evaluation

To evaluate how well our model was able to predict future news-related activity using features derived from the news cycle (i.e., the number and sentiment polarity of articles published on the target day and over a rolling window), we trained our model on the first $n\%$ of days for each event and tested the model on the remaining $(100 - n)\%$ of days. We used NRMSE (normalized root-mean-square error) to calculate the normalized difference between the number of news-related tweets we predicted and the number of news-related tweets in our ground truth.

Our results as n varies are shown in Table 2. Using only the activity from the first 20% of an event, we are able to estimate the amount of news-related tweets with a NRMSE below 0.2 for most events. The events' NRMSE's are strongly correlated with the events' metadata when $n = 30\%$. Spearman's ρ between the NRMSEs and total number of tweets is

0.82; similarly, the correlation between the NRMSEs and the total number of news articles is 0.61. These correlations do not hold true when less training data is used, however, with respective correlations of 0.35 and 0.09 when $n = 20\%$. Both correlations are similarly low when $n = 10\%$. Both correlations are strong when $n = 40\%$ and $n = 50\%$, as they were when $n = 30\%$.

The NRMSEs decline further with more data, but the value of predicting activity may also decline as an event continues. The mean NRMSE across all events is shown in Figure 1. The mean NRMSE drops below 0.15 once 30% of the data is used and continues to decline as additional training data is added.

Figures 2 and 3 illustrate the model's predictions on the BP Oil Spill event when moving from using 30% to 40% of the event's training data, respectively. Solid lines indicate actual tweet counts and the dotted lines indicate the model's predictions. In both cases the model follows the trend of the event, but fails to correctly estimate the magnitude of the increase in activity in late May. The models' predictions for other time periods are reasonably accurate, however, and in both cases the NRMSE remains under 0.20. The model's tweet underestimation can be partially attributed to the fact that news article counts are always relatively low, which makes it difficult to estimate how high a spike in Twitter activity should be. In the case of the BP Oil Spill, the model performed poorly as new developments continued to unfold in late May (e.g., commentators suggested the extent of the spill was greater than previously believed and discussed methods for plugging the well). Such developments could be captured by features designed to measure the novelty of new news articles, but we leave incorporating such features into our model as future work.

Conclusion

We described "news-centric events" (i.e., events for which the vast majority of tweets are posted in response to news articles) and proposed a method for predicting the number of tweets posted about such events (i.e., "news-related tweets") based only on the news articles published about the event. Our model's accuracy shows that, for news-centric events, a relationship exists between the tweets and news articles written about an event. We leave as future work the questions of whether this relationship exists for other types of events and whether our model can be extended to predict the number of news-related tweets about non-news-centric events.

Acknowledgments

This work was partially supported by the National Science Foundation through grant CNS-1204347 and through REU award IIP-1362046.

References

Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S. H.; and Liu, B. 2011. Predicting flu trends using twitter data. In *2011 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2011*, 702–707.

Asur, S., and Huberman, B. a. 2010. Predicting the Future with Social Media. *Computing* 1:492–499.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* 0:2200–2204.

Cooper, C. P.; Mallon, K. P.; Leadbetter, S.; Pollack, L. a.; and Peipins, L. a. 2005. Cancer internet search activity on a major search engine, United States 2001-2003. *Journal of Medical Internet Research* 7(3):1–13.

Copeland, P.; Romano, R.; Zhang, T.; Hecht, G.; Zigmond, D.; and Stefansen, C. 2013. Google Disease Trends: An update. *Nature* 457(Cdc):1012—-1014.

Dugas, A. F.; Jalalpour, M.; Gel, Y.; Levin, S.; Torcaso, F.; Igusa, T.; and Rothman, R. E. 2013. Influenza Forecasting with Google Flu Trends. *PLoS ONE* 8(1):2579.

Glasgow, K.; Fink, C.; and Boyd-Graber, J. 2014. Our grief is unspeakable: Automatically measuring the community impact of a tragedy. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM'14.

Goel, S.; Hofman, J. M.; Lahaie, S.; Pennock, D. M.; and Watts, D. J. 2010. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America* 107(41):17486–17490.

Mishne, G., and Glance, N. 2005. Predicting Movie Sales from Blogger Sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 155–158.

O'Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM'10.

Oghina, A.; Breuss, M.; Tsagkias, M.; and de Rijke, M. 2012. Predicting imdb movie ratings using social media. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, 503–507.

Paul, M., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM'11.

Pearson Education. 2015. Information Please: News and Events. http://www.infoplease.com/yearbyyear.html.

Polgreen, P. M.; Chen, Y.; Pennock, D. M.; and Nelson, F. D. 2008. Using internet searches for influenza surveillance. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 47:1443–1448.

Signorini, A.; Segre, A. M.; and Polgreen, P. M. 2011. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*.

Tumasjan, A.; Sprenger, T.; Sandner, P.; and Welpe, I. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM'10.