Does Bad News Go Away Faster? *

Shaomei Wu and Chenhao Tan and Jon Kleinberg and Michael Macy
Cornell University
Ithaca, New York 14850

Abstract

We study the relationship between content and temporal dynamics of information on Twitter, focusing on the *persistence* of information. We compare two extreme temporal patterns in the decay rate of URLs embedded in tweets, defining a prediction task to distinguish between URLs that fade rapidly following their peak of popularity and those that fade more slowly. Our experiments show a strong association between the content and the temporal dynamics of information: given unigram features extracted from corresponding HTML webpages, a linear SVM classifier can predict the temporal pattern of URLs with high accuracy. We further explore the content of URLs in the two temporal classes using various textual analysis techniques (via LIWC and trend detection). We find that the rapidly-fading information contains significantly more words related to negative emotion, actions, and more complicated cognitive processes, whereas the persistent information contains more words related to positive emotion, leisure, and lifestyle.

Introduction

Several previous studies have revealed distinctive temporal patterns of information dissemination in various social media domains (Gruhl et al. 2004; Leskovec, Backstrom, and Kleinberg 2009; Crane and Sornette 2008; Yang and Leskovec 2011; Wu et al. 2011). Other studies have sought to identify the underlying mechanisms that drive the temporal dynamics, including recent papers that have attempted to predict which pieces of content will produce large spikes of attention, or large cascades in social networks (Yang and Leskovec 2011; Hansen et al. 2011; Bakshy et al. 2011). However, even for content that achieves wide circulation, the "shape" of the interest in this content over time can vary considerably (Leskovec, Backstrom, and Kleinberg 2009; Yang and Leskovec 2011). Here, we formulate a novel prediction task focusing on the the *persistence* of a piece of online information, as measured by how long it continues to generate attention after its peak.

*Supported in part by the MacArthur Foundation, a Google Research Grant, a Yahoo! Research Alliance Grant, a grant from Microsoft, ONR (YIP-N000140910911), and NSF grants IIS-0910664, CCF-0910940, IIS-1016099, and DMS-0808864.

Using public data from Twitter, we track the flow and persistence of attention to URLs embedded in tweets. Our goal is to look for intrinsic qualities of the content that influence the persistence of information. Our paper makes three main contributions:

- We build a classifier that predicts the decay/persistence of information with textual features, providing one of the first empirical studies of the connection between content and temporal variations of information in social media.
- We investigate the properties of the text that are associated with different temporal patterns, finding significant differences in word usage and sentiment between rapidly-fading and long-lasting information.
- By measuring the temporal pattern of information based on content alone, we are able to predict the long-term trajectory at a very early stage, when the information is first generated, which may be useful for social movement activists, public relations agencies, and advertisers.

Data

We used a dataset publicly shared by the authors of (Yang and Leskovec 2011)[1], consisting of approximately 20%-30% of all the tweets generated between June 1, 2009 and December 31, 2009. We only study the temporal patterns of URLs as they are easily identifiable and represent a much richer source of content beyond the 140-character limit of tweets. From the total 476M tweets contained in the dataset, we find 118M distinct URLs embedded in 186M tweets. Among all the URLs, nearly half (56M) are bit.ly URLs (i.e., start with http://bit.ly/). For simplicity, we only extract the time series of bit.ly URLs and use them as a representative sample of all temporal patterns. We further restrict our study to URLs that are mentioned more than 50 times in total and more than 10 times in retweets[2], in order to remove spam and have sufficient observations to measure temporal dynamics, leaving 24K URLs. Of these, we were able to crawl 21K; the remaining 3K were either misspelled or linked to web pages that no longer exist. Thus, our analysis is limited to the temporal pattern in the time series of these 21K bit.ly URLs.

[1] http://snap.stanford.edu/data/twitter7.html

[2] We recognize a post as retweet when it contains "RT @" or "via @".

Persistence of URLs

We measure the persistence of content using the decay rates following peak attention rather than first occurrence. First occurrence can be followed by long gaps, which confounds persistence with slow acceleration in the rate of attention.

For each URL u, let the hour of maximum attention (also called the peak of attention) be hour 0. Then the *decay time* t_u is defined as the hour after the peak when the number of mentions first reaches 75% of the total. We intentionally choose to measure the time lag from the peak of attention to the point when the URL fades away, to reduce the possible censoring bias given the limited observation window when the dataset was collected (Wu et al. 2011). The distribution of t_u approximately follows a power-law (see Figure 1), as found previously in the distribution of URL lifespan (Wu et al. 2011). Among all URLs we studied, the mean t_u is 217.3 hours and the median t_u is 19 hours.

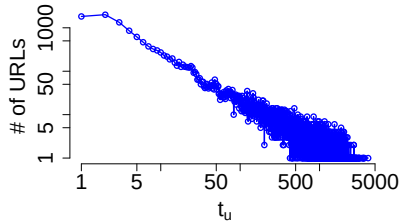

Figure 1: Distribution of URL decay time t_u

Predicting temporal patterns based on content

In this section, we formally define the temporal pattern classification task and present our findings.

Identifying two distinct temporal patterns

We begin by comparing the content of rapidly-fading and long-lasting URLs, using a binary classifier. Previous studies have found 24-hours to be a typical news cycle and content that lasts more than 24 hours usually attracts consistent waves of attention for a long period of time (Leskovec, Backstrom, and Kleinberg 2009; Yang and Leskovec 2011). We thus define class 1 as consisting of those URLs with $t_u > 24$. In this way, we get a positive class of persistent content with 7042 examples. We define class 0 as consisting of those URLs with $t_u < 6$, which gives us 6185 examples. We choose the cutoff value 6 here to get a balanced distribution of positive and negative examples. Our definition of two classes makes the prediction problem much more tractable, and also provides us enough examples to achieve reliable estimates.

To better illustrate our classification scheme, we apply the time series normalization method introduced in (Yang and Leskovec 2011) and calculate the centroid of the time series for each class, as shown in Figure 2. There is a clear difference between the normalized temporal pattern of the two classes: URLs of the positive class fade away slowly, with periodic, multiple peaks of attention; URLs of the negative class have a single spike and a rapid decay afterwards.

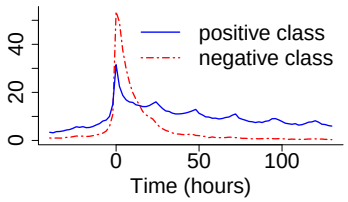

Figure 2: Normalized time series centroids for two classes

Table 1: Results for predicting lastingness of information

Feature	*Accuracy*	*Pos F1*	*Neg F1*
Header	0.6909	0.7399	0.6186
Header + URL	0.7177	0.7666	0.6423
Header + Body	0.7136	0.7664	0.6296
Header + Body + URL	0.7224	0.7708	0.6478

Features

To predict the temporal class of URLs, we extract the following four incremental sets of unigram features from the HTML webpages linked by the URLs:

- Header. The text in the header of HTML, within tags "<title>","<description>", and "<keywords>".
- Header + URL. In addition to Header, this feature set also uses the terms tokenized from the URL links embedded in the HTML (i.e.,within "<href>").
- Header + Body. In addition to Header, this feature set includes all the text in the body of HTML.
- Header + URL + Body. This feature set combines all the features mentioned above.

To get more meaningful unigram features, after tokenizing all the textual content into word terms, we filter the terms with length 1 (e.g., "s", "t") and the terms consisting of only numbers. As the dimension of the feature space increases sharply in the last 3 sets of features, we also filter the infrequent terms (i.e., terms with total frequency less than 20). In this way, we get 18471 unique unigram terms in Header, 27433 in Header + URL, 59475 in Header + Body, and 76487 in Header + Body + URL.

Classifier performance

To predict the persistence of webpages, we employ a Support Vector Machine (SVM)[3] classifier with a binary representation of unigram features (if a term appears in a webpage, the corresponding coordinate has value 1, and value 0 otherwise). To work with high-dimensional features, we use the linear SVM kernel for efficiency. We also apply the default parameters for the SVM classifier for a fair comparison among different sets of features. Table 1 gives the performance of classifiers with different sets of features using 10-fold cross validation.

Table 1 shows that in general, the simple linear-kernel SVM classifier can predict the temporal category of URLs

[3]SVMLight: `http://svmlight.joachims.org/`

with impressively high accuracy (around 70%), as compared to 53% for always predicting positive. Also, the F1 score for positive class is around 75%, which shows a remarkable balance of precision and recall at identifying the persistent content. This result provides strong evidence for the connection between the content of HTML pages and the persistence of the associated URLs. Comparing across 4 feature sets, we see that the more information we have about the content, the better the classifier performs. This finding further confirms the relationship between textual content and the persistence of attention to the information.

How temporal patterns vary with content

The SVM classifier shows that the content provides sufficient information to predict the persistence of information. However, SVMs are not as effective at identifying meaningful properties of the text that are most related to the differences in temporal patterns. We address this problem by examining the text with easily interpretable content-analysis methods that identify content that exhibits the largest difference across temporal classes.

LIWC analysis

Linguistic Inquiry and Word Count (LIWC)[4] is a widely used text analysis tool that maps words into 60 pre-defined categories, covering linguistic, psychological, and social dimensions. We start by comparing the distribution of LIWC categories across two classes.

We say a LIWC category occurs in a URL when we find at least one word under that category from the header of the associated HTML page.[5] Figure 3 shows the percentage of occurrence for all LIWC categories in webpages from the two classes. The two classes differ the most in the following three groups of LIWC categories (see Figure 3),

- Emotion: *posemo* (positive emotion), *negemo* (negative emotion).
- Cognitive process: *cogmech* (cognitive process), *insight* (words like *think*, *know*, *consider*), *incl* (inclusive, words like *and*, *with*, *include*), *discrep* (discrepancy, words like *should*, *would*, *count*).
- Part of speech: *verb* (common verbs), *auxverb* (auxiliary verbs), *preps* (prepositions), *present* (present tense, words like *is,does*, *hear*), *future* (future tense, words like *will,gonna*).

To better see the trend in the frequency of specific categories as a function of t_u, for each category w, we define $f_w(t)$ as the fraction of occurrences of w in all URLs u for which $t_u = t$, and plot $y = f_w(t)$ for different groups of LIWC categories in Figure 4.

[4] `http://www.liwc.net/`

[5] We also conduct the same analysis with text from the other 3 feature sets, however, since the number of words increases markedly in these feature sets, and the LIWC dictionary many times maps a word into multiple categories, the binary vector for each URL is easily saturated and the $f_w(t)$ curve becomes too flat to show interesting differences.

Table 2: Representative words for two temporal classes

class	*representative words*
pos	twibbon, marketing, contest, trailer, review, support, vote, giveaway, big, movie, design, quot, win, good, best, love, green, week, funny, version
neg	cnn, blogs, source, finest, onion, apple, house, iphone, white, guardian, google, users, app, download, america, jackson, public, myspace, today, uk

Again, to balance the power-law distribution of t_u, we bin t_u by the integer part of $\log_2(t_u)$, and plot the value $f_x(w)$ for each bin x (instead of hour x). Thus, the later bins would still contain a substantial number of URLs so that the probabilistic curve is smoother.

Similar to previous studies (Berger and Milkman 2010; Hansen et al. 2011), we find that affective content of a URL is related to its persistence. URLs containing words with positive emotion are more persistent than those with negative emotion. However, the number of words related to affect remains more or less constant across t_u. We also see a drop in the number of words related to cogntive process as t_u increases, suggesting that content associated with more complicated cognitive processes can be more viral (Berger and Milkman 2010), yet not as persistent. We find that rapidly-fading URLs point to content with more words related to actions (verb, auxverb, preps) and tense (present, future), potentially because these webpages contain more action-demanding, time-critical information that expires after a certain event or time.

Trending words analysis

As a manually-generated pre-defined category system, LIWC is limited by the underlying psycholinguistic concepts. To extend the dimensions of text described in LIWC, we apply the trend detection techniques with relative change metrics (Kleinberg 2004) and compare the top 20 most representative header words for the two classes.[6]

To identify the words that are most meaningful, we filter out numbers and all words with frequency less than 20 (mostly specific names) or greater than 400 (mostly stopwords and website names). The results in Table 2 reinforce and provide intuitive interpretation for the LIWC results. We again find the persistent URLS are more likely to point to text containing positive words (e.g. good, best, love). In terms of the semantics of content, the persistent webpages are more related to art (e.g. music, movie), advertisement, and online marketing (e.g. twibbon, marketing, giveaway, free, win, review), whereas the rapidly-fading webpages contain more news (e.g. cnn, google, onion, guardian, blogs), and names (e.g. michael jackson, white house, obama, iran, america, uk). The association between news and short-lived URLs and between art and long-lasting URLs supports previous findings (Wu et al. 2011).

[6] For the other three feature sets, as the number of terms increases, the data becomes too noisy to be described with a few words, the results thus are difficult to interpret.

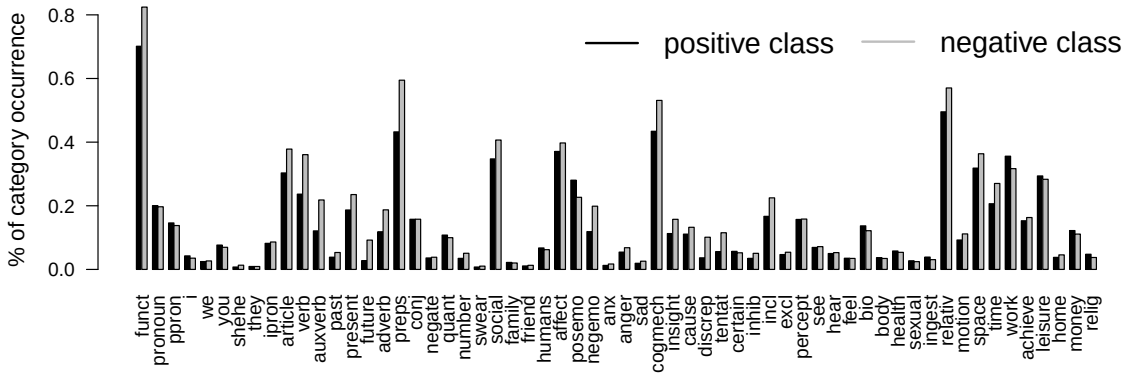

Figure 3: Class distribution in 60 LIWC dimensions, using words from HTML header

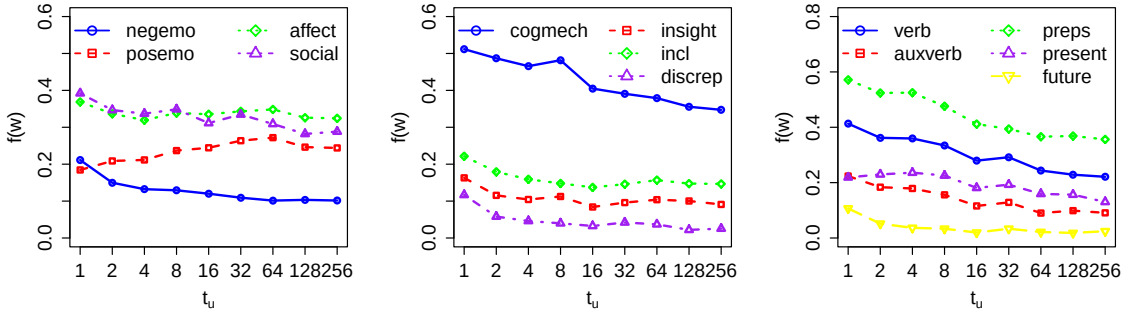

Figure 4: Trending LIWC categories

Conclusion and future work

We have explored the relationship between the content and the persistence of information as measured by decay time, in the context of Twitter. We find that by using the textual features extracted from the content, we can predict the persistence of information with high accuracy.

We also use different text analysis techniques to identify content that contributes to persistence. To that end, we compared psycholinguistic characteristics, and trending words in content pointed to by rapidly-fading and long-lasting URLs. Results show that persistent information tends to express positive affect and refer to art, while rapidly-fading content tends to contain time-critical information (e.g., news) that carries relatively more negative sentiments, demands more cognitive effort, or is associated with quick action.

This study of the time series of bit.ly URLs posted on Twitter may limit our findings to social media and the dynamics of information they support. Assessing the broader implications will require future work that extends the analysis to other types of information and systems of communication, such as the transcribed content of TV, radio, and print media.

Finally, we only predict the persistence of content for two extreme cases. Future studies might usefully investigate the connection between content and persistence across the full range of outcomes, including information that is neither rapidly-fading nor long-lasting.

References

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *WSDM '11*.

Berger, J., and Milkman, K. 2010. Social transmission, emotion, and the virality of online content. *Wharton Research Paper*.

Crane, R., and Sornette, D. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* 105(41):15649–15653.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW'04*.

Hansen, L. K.; Arvidsson, A.; Nielsen, F. Å.; Colleoni, E.; and Etter, M. 2011. Good friends, bad news - affect and virality in twitter. *CoRR* abs/1101.0510.

Kleinberg, J. 2004. Temporal dynamics of on-line information streams. In *Data Stream Managemnt: Processing High-speed Data*. Springer.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD'09*, 497–506.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on twitter. In *WWW '11*.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *WSDM'11*.