Does Bad News Go Away Faster?

Shaomei Wu
Department of Information Science, Cornell University

Chenhao Tan, Jon Kleinberg
Department of Computer Science, Cornell University

Michael Macy
Department of Sociology, Cornell University

Introduction

Several previous studies have revealed distinctive temporal patterns of information dissemination in various social media domains [2,3,5,6,7] (see Figure 1 for examples). Here, we study the relationship between content and temporal dynamics of information on Twitter, focusing on the *persistence* of information. Using public data from Twitter, we track the occurrences of URLs embedded in tweets, and measure the persistence of a URL (a unit of online information) by how long it continues to appear after its peak.

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-tem trajectory at a very early stage, when the information is first generated**.

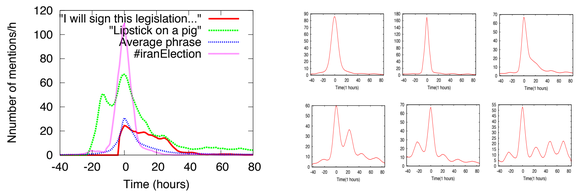

Figure 1.Examples of temporal pattern of to information dissemination on social media [7]

Data

We use the same dataset as used in [7], consisting of 476M tweets generated between 6/1/2009 to 12/31/2009. In this dataset, we find 118M distinct URLs embedded in 186M tweets. Since half of them (56M) are bit.ly URLs, 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, in order to remove spam and have sufficient observations to measure the temporal dynamics, leaving 24K URLs. Of these, we were able to crawl 21K (the rest are either misspelled or linked to pages that no longer exist). Thus, our analysis is limited to the temporal pattern of the occurrences of these 21K bit.ly URLs.

Persistence of URLs

We measure the persistence of content using the decay rates following peak attention. For each URL u, let the hour of peak 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. The distribution of t_u is shown in Figure 2. Among all the URLs we studied, the mean t_u is *217.3* hours and the median t_u is *19* hours.

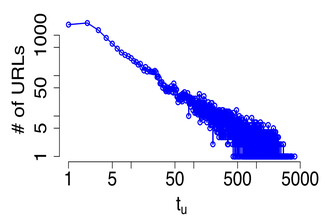

Figure 2. Distribution of URL decay time t_u

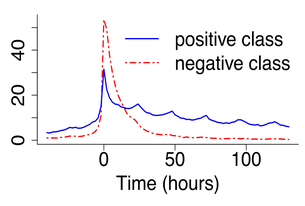

Figure 3. Normalized time series centroids for two classes

Predicting Temporal Patterns Based on Content

We begin by comparing the content of *rapidly-fading* and *long-lasting* URLs, using a binary classifier.

Identifying Two Distinct Temporal Patterns

Previous studies found 24-hours to be a typical news cycle and content that lasts more than 24 hours usually attracts consistent waves of attention [5,7], we thus define class 1 as consisting of those URLs with $t_u > 24$, and get a positive class of persistent content with *7042* examples. To balance the distribution of positive and negative example, we define class 0 by those URLs with $t_u < 6$, which gives us *6185* examples. We apply the time series normalization method introduced in [x], and calculate the centroid of the time series for each class (see Figure 3).

Features

We extract the following four incremental sets of unigram features from the HTML webpages linked by the URLs:

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

Note that we filter the terms with length 1, the terms consisting of only numbers, and the infrequent terms (i.e., terms that occur less than 20 times).

Classifier Performance

To predict the persistence of webpages, we use a SVM classifier with a binary representation of unigram features (if a term appears in a webpage, the corresponding coordinate has value 1, otherwise 0). We use the linear kernel for efficiency. Table 1 gives the performance of classifiers with different sets of features using 10-fold cross validation.

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

Table 1 shows that the simple linear-kernal SVM classifier can predict the temporal category of URLs with impressively high accuracy and a good balance of precision and recall (for identifying persistent content). This result provides strong evidence for the connection between the content and the persistence of attention to the information. Comparing across 4 feature sets, we see that the more information we have about the content, the better the classifier performs.

How Temporal Patterns Vary with Content

As SVMs are not as effective at identifying meaningful properties of the text that are most related to the differences in temporal patterns. In this part, we examine the text with easily interpretable content-analysis methods that identify characteristics of content that exhibits the largest difference across temporal classes.

LIWC Analysis

Linguistic Inquiry and Word Count (LIWC) is a widely used text analysis tool that maps words into 60 pre-defined categories, in linguistic, psychological, and social dimensions.

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. The distribution of LIWC categories across two classes is shown in Figure 4.

To show 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 (see Figure 5). We bin t_u by the integer part of $\log_2(t_u)$, and plot the value of $f_x(w)$ for each bin x.

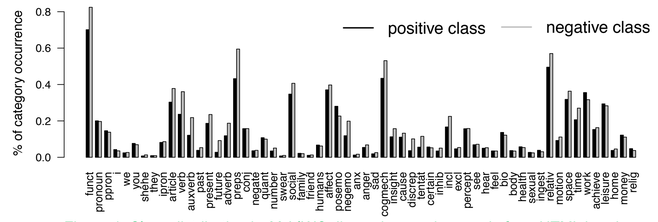

Figure 4. Class distribution in 60 LIWC dimensions, using words from HTML header

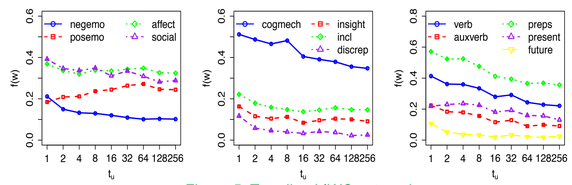

Figure 5. Trending LIWC categories

Figure 4 shows that the two classes differ the most in the three groups of categories:

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

In Figure 5, we see that:

- The amount affect words remains constant across tu, but positive emotion are more persistent than negative emotion;
- Content associated to more complicated cognitive process is not very persistent;
- Rapidly-fading content has more words related to actions (verb, auxverb, preps) and tense (present, future).

Trending Words Analysis

To extend the dimensions of text described in LIWC, we apply the trend detection techniques [4] and compare the top 20 most representative header words for the two classes (see Table 2).

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

The results in Table 2 provide intuitive interpretation for the LIWC results. We find that:

- Persistent URLs are more likely to point to text containing positive words;
- Persistent webpages are more related to art, advertisement, and online marketing;
- Rapidly-fading webpages contain more news and names.

Conclusion

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 compare psycholinguistic characteristics, and trending words in content of rapid-fading and long-lasting categories. We find that persistent information tends to express positive affect and refer to art topics, while rapidly-fading content tends to contain time-critical information that carries relatively more negative sentiments, demands more cognitive effort, or is associated with quick action.

References

[1] Berger, J., and Milkman, K. 2010. Social transmission, emotion, and the virality of online content. Wharton Research.
[2] 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.
[3] Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In WWW'04.
[4] Kleinberg, J. 2004. Temporal dynamics of on-line information streams. In Data Stream Managemnt: Processing High-speed Data. Springer.
[5] Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Memetracking and the dynamics of the news cycle. In KDD'09.
[6] Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on twitter. In WWW '11.
[7] Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In WSDM'11.