UNIVERSITEIT VAN AMSTERDAM

Predicting the Propagation of Video Content on Twitter

Philo van KEMENADE
5894875
philovankemenade [at] gmail.com

Thesis
BSc. Artificial Intelligence
Credits: 15

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
M. Worring

Intelligent Systems Lab Amsterdam
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

July 23, 2011

Abstract

Online social networks have become widespread and increasingly popular. A common activity on these networks is sharing of content. Since the advent of social media hosting sites, users have vast repositories of external content to share. For content creators interested in the impact of their work, prediction of the propagation of their content is a valuable asset. In this thesis we work on predicting the spread of online video content on the social network Twitter. We take two perspectives in extracting information about videos and how they are shared. A feature-based model identifies the content and meta data of a video. Our propagation model describes how the sharing of a video develops over time. By indicating correlations between features of the video content on one side and characteristics of propagation patterns on the other, we show how videos spread differently depending on their category, length and age. We use these findings to form a non-linear prediction model, that forecasts social propagation of a video based on it's features and an initial propagation period. Our evaluations show that using our description models in combination with extended initial periods, can yield results of predicting slightly under 90% of the instances in our experiment within a relative error margin of 20%.

Keywords: social media, online video, content sharing, information propagation, prediction

Contents

1 Introduction

Over the past few years, social networking sites have become very popular and are used by an increasing number of people. People use these networks for many different reasons, but an important activity is sharing of content [10]. Often, social networks are a fundamental medium for the spread of content people choose to share. When people share information on a network, others may decide to further propagate messages they receive. By this kind of person to person propagation of information, over time information cascades are formed that allow information to spread far and wide across a network [21]. For people who care about the impact of their work, such as artists, marketeers or politicians it is a useful asset to know how their content is propagated online. Moreover, it is extremely valuable for them to be able to have grounded expectations about the future development of such propagations. In this thesis, we work on this problem of predicting future propagation of content across a social network.

We focus our research specifically on the propagation of online video content, which is shared extensively on social networks [12]. We are interested to see whether videos that differ in content, also have a different propagation structure. If so, it makes sense to base predictions of future propagation in part on features of the video that is spread. The propagation structures that we study in this work are those caused by the propagation of videos on the microblogging service Twitter. On this large social network, people communicate in short, 140 character long 'tweets'. People connect to each other by 'following' someone to thereby receive updates of the tweets this user sends out.

We take two perspectives in describing videos shared on Twitter. On one side we consider a video's content features that convey information about the type of video and its metadata such as length and age. On the other side we investigate how a video propagates over the Twitter network and extract features to build a description model of its propagation. We first use this two-sided description of videos to show correlations between content and prediction. Next we build prediction models based on coupled features from both sides to predict four measures of future propagation; tweets, retweets, users and followers.

The rest of the paper is organized as follows. We will sketch the context of our research and report on related work in section 2, we give an overview of our two-sided perspective to videos and their propagation in section 3. Next we present our methodology in section 4 and describe our experimental setup in section 5. We discuss the implications of our results and conclude in section 6.

2 Related Work

In this section we report on related work in fields dealing with prediction of network dynamics, online content sharing and diffusion of information in social networks.

The topic of prediction of social network dynamics has seen attention from different perspectives. In [16] Nowell and Kleinberg predict the formation of new internal links in a social network. More on a information propagation note, various works have worked on influence maximization, the problem of selecting an optimal subset of users in a social network to aid the diffusion of information [11, 4]. [20] hints at using findings about URL and tag dynamics on Twitter for prediction of future propagation of URLs. Yang and Counts have studied information diffusion on Twitter and found that properties of users and their interactions are telling for the prediction of local propagation structures [22].

Content sharing behavior on the web has been analyzed in a number of previous studies. For example Kinsela et al. have studied the practice of posting links in online conversations over a period of ten years. Their work shows that people not only link more to external sites, the URLs nowadays often link to visual media content, of which videos make up a large part [12]. In [3] Cha, Prez and Haddadi analyze structural properties of the blogosphere next to content sharing patterns on blogs. This latter part is of particular interest for our present topic as they indicate important correlations between video content and propagation patterns on blogs.

Analysis of the spread of information in networks has been around for a long time [19, 7] and is applied to different fields like viral marketing [11], epidemiology [18] and social sciences [7, 14, 8]. Recently, with the increased popularity of online social and information networks, the topic has seen much attention from computer science research [13]. Altogether, different approaches are taken. Various studies have analyzed networks on a node to node level, looking at the topology of social networks and thus enabling discussion of local characteristics and graph structures [15, 5].

Another approach is also possible, where characteristics of diffusion are analyzed on a more global level, without looking at individual social ties within a network. Both approaches are often combined [2, 3, 6, 13]. A good example of the more global approach can be seen in the work of Cha, Mislove and Gummadi who analyze temporal evolution of picture popularity [2]. They analyze growth patterns over time by measuring cumulative fan growth of large collections of photographs. From a social science perspective, Kwak et al. study the medium Twitter and pay attention to the development of trending topics. They report on different types of tweets, user participation and active periods of trends [14].

In this thesis we focus in part on the topic of indicating correlations between video content and video propagation. However, we extend the analysis in [3] both from the perspective of video content and that of propagation patterns. Apart from indicating general patterns in video propagation, we show how to use this knowledge for prediction. Our prediction models achieve promising results that not only indicate which features are useful for the prediction of propagation, but also what level of performance can be reached.

3 One Video, Two Perspectives

This section sketches our approach to the propagation of video content. We indicate the perspectives we take towards video content and the way it propagates on a social network.

On the open communication platform Twitter where message size is limited to 140 characters, a convenient way of sharing information is linking to external content by including HTML URLs in a message. We are interested in tweets containing a link to video material, to see how specific videos spread over the network. Given a single video, we take two perspectives in constructing its description. One describes a video's characteristics in terms of content features by looking at its source. The other describes the behavior of the video, or in fact that of its viewers, on a social network. A sketch of this approach is included in figure 1. By coupling these two descriptions of a video and its propagation, we come to a coherent description model to use in our task of predicting future propagation.

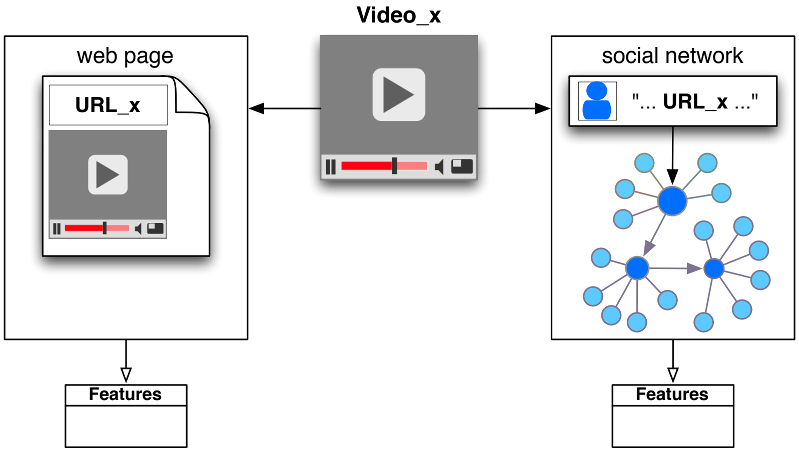


Figure 1: An overview of our two perspectives to a video with their respective features

3.1 Video Source

To obtain information about a video's content, we look at the web page where a video is hosted. In order to identify a video's characteristics, various sources may be used. The most direct way to get a description of a video, would be to analyze the video material itself and extract semantic features about the topic, entities and featured content in the video. Although this method might lead to a rich set of content descriptors, it is computationally expensive if done automatically and strenuously time-consuming otherwise.

Another, more computationally friendly source of information about a video, is its metadata as provided by the hosting website. The last years have seen an enormous increase in the amount of video being uploaded to large hosting sites like Youtube [23]. Moreover, Kinsela et al. indicate a related trend of increased linkage in online conversations to content on websites for which there is structured data available. By accessing a webpage's source to parse metadata, features such as video length, upload date, and video category can be extracted. As we will review in section 4, we will take this second approach to build a description model for the content of videos.

3.2 Propagation on Network

Once uploaded, viewers of a video may decide to share it with their peers on a social network. Doing so can cause an information cascade to be formed when other users decide to propagate the shared information further to their respective connections. When this is done extensively, vast areas of the network can be reached. Looking at this on a global level, large scale diffusion of information can be established [19].

We take a standpoint on this global level and look how properties of the set of all tweets concerning a specific video, develop over time. We call such a collection of tweets and its development the *propagation* of a video. By looking at how propagations evolve from day to day, we get information about the overall spread of videos. This aggregate approach allows us to work with data that is sampled uniformly over time. Sampled data prohibits a graph-based modeling approach for aspects like retweet structure or the cascade of a single piece of information, as has been used in other works [20, 1, 14]. We will not regard these issues, but focus solely on the aggregate-level aspects of propagation. Before presenting these different aspects, we first introduce our terminology.

We define the Twitter network as the directed graph $G = (V, E)$, where each $v \in V$ is a user. For any two users $v_i \in V$, $v_j \in V$ there is a directed edge $(v_i, v_j) \in E$ if v_j is a follower of v_i. In other words, edges run from user v_i to another user v_j, when information flows from v_i to v_j due to v_j's subscriptions to tweets of v_i. Reversely;

Definition 1. *A user $v_j \in V$ is a follower of $v_i \in V$ if there exists an edge $(v_i, v_j) \in E$*

We further employ the following terminology:

Definition 2. *A tweet is a triple $< v, m, t >$, where $v \in V$ is the user who posted the tweet, m is the text (message) of the tweet and t is the time at which the tweet is posted.*

Definition 3. *For a given URL x, we say that a user $v \in V$ mentions x if there exists a tweet $< v, m, t >$ that contains x in m.*

Definition 4. *For a given URL x, a user $v_j \in V$ is reached by x if there is a user $v_i \in V$ that mentions x and v_j is a follower of v_i*

Definition 5. *A tweet $< v_2, m_2, t_2 >$ is a retweet of tweet $< v_1, m_1, t_1 >$ if m_2 includes "RT@v_1" and a part of m_1. If $< v_2, m_2, t_2 >$ is a retweet of $< v_1, m_1, t_1 >$, it follows naturally that $t_2 > t_1$*

Definition 6. *Given a video c (content) with associated URL x, the propagation P_c of c is the set of all tweets $< v, m, t >$ for which m contains x.*

To study a video's propagation we extract information about different aspects of the communication of a video and measure their development over time. We use measures that concern persons involved in the spread of a video as well as the messages created by these persons. We discuss these methods specifically in section 4.2.

4 Methodology

This section presents our methods to build our two-sided description model of videos and their propagation. We explain how we inspect correlations between the two types of data and how we then use this for prediction.

4.1 Features from Video Source

We obtain features of a video from its metadata as provided by its web source. Given a single video, we parse its HTML source to extract the following features:

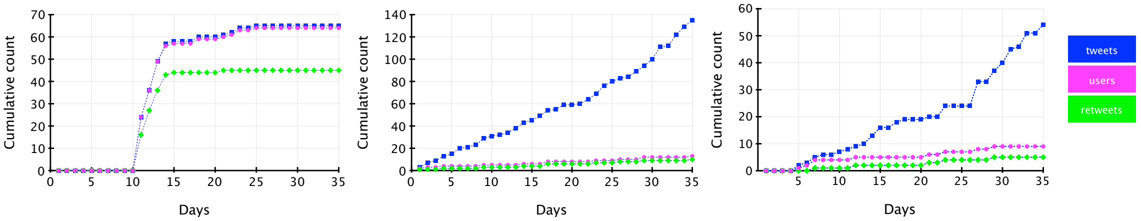
- *Video Title*
- *Upload Date*
- *Video Length*
- *Video Category*; uploaders can choose one of 15 specified categories
- *HD status*; Whether or not HD resolution is available

4.2 Propagation model

Given a period in time, we collect the set of tweets that contain URL x of video c, and track four metrics in the propagation P_c:

- *Tweets*; the number of tweets $< v, m, t > \in P_c$
- *Retweets*; the number of tweets $< v, m, t > \in P_c$, such that $< v, m, t >$ is a retweet
- *Users*; the number of unique users $v \in V$ for which there exists a tweet $< v, m, t > \in P_c$
- *Followers*; the number of users $v_j \in V$, for which there exists an edge $(v_i, v_j) \in E$ from another user $v_i \in V$ and v_i mentions x

Given a video and its corresponding URL, we first sort all tweets belonging to the video's propagation on post time and measure number of tweets, retweets, users and followers per day. Figure 2 shows two examples of such propagations. For sake of readability, we have excluded followers as these numbers exceed the other measures by some orders of magnitude. Their development shows the same structure as that of tweets, retweets and users; a logarithmic-shaped plot for the video of figure 2a and a more linear, though rippled plot for the video of figures 2b and 2c.



(a) (b) (c)

Figure 2: Examples of video propagations

For each of the four aspects we extract features that capture properties of their evolution over time. A first basic feature is the total count that is acquired during a given period in time:

$$total\ count = \sum_{k=1}^{\#days} count_k$$

Spreading Time

A first notion we capture in the development of the four quantities mentioned above, is the amount of time a video takes to spread. In [3], Cha et al. propose two measures for the duration of a propagation. *Full-spreading time* is the number of days of the entire propagation, from the first to the last post of the URL of a given video:

$$\textit{Full-spreading time} = t_{max} - t_{min}$$
$$\textit{where } t_{max}, t_{min} \in P_c$$

Half-spreading time is the number of days that the propagation of a URL takes, starting from the first post, until 50% of all mentions have appeared:

$$\textit{Half-spreading time} = t_{half} - t_{min}$$
$$\textit{where } t_{half}, t_{min} \in P_c \textit{ and } t_{half} \textit{ such that } \sum_{k=1}^{t_{half}} count_k = 0.5 * \textit{total count}$$

While *full-spreading time* simply measures the duration of a propagation, *half-spreading time* informs more detailed about the initial speed of the propagation.

Looking at the examples in figure 2, we see how both measures differ. For the number of tweets for example, the video of figure 2a only starts to spread after day 10 and thus has a *full-spreading time* of 25 days, whereas the video of figure 2b is actively spread over the complete period of time and thus has a *full-spreading time* of 35 days. Also the *half-spreading time* differs. Because of the relatively quick increase during the first four days of propagation in figure 2a, the *half-spreading time* for tweets is two days here, while being 23 days in figure 2b.

As these figures show, information about half and full spreading times gives some differentiating power in describing propagation structure. However, the relation between the two tells a lot more about the propagation than these two numbers as mere separate quantities. To account for the relation between the two we introduce a feature for the ratio of *half-spreading time* and *full-spreading time*:

$$\textit{half-spread ratio} = \frac{\textit{half-spreading time}}{\textit{full-spreading time}}$$

Propagations with a logarithmic-shaped curve like the one in figure 2b generally have a low *half-spread ratio* (0.08 here), that indicates a relatively fast initial spread followed by a decrease in propagation. Propagations with a more linear curve like the one in figure 2a have a *half-spread ratio* around 0.5.

Spreading Speed

Another part of a propagation's structure we take into account, is the speed at which a video spreads. Consider the propagation in figures 2b and 2c. Both show a linear development and their respective *half-spread ratios* of 0.66 and 0.74 are fairly similar. However, a far lesser number of tweets is produced during the propagation in figure 2b than during that in figure 2c. We consider such differences in the gradient of propagations by introducing a feature for spreading speed. We specifically measure the speed of the second half of the propagation, as this is most telling for the future development we eventually want to predict:

$$\textit{spread speed} = \frac{0.5 * \ \textit{total count}}{\textit{full-spreading time} - \textit{half-spreading time}}$$

This feature allows for better differentiating between the two figures 2b and 2c, with their respective values for tweets reading 5.5 and 2.75 tweets per day.

Retweets

While some propagations, such as the one in figure 2b, almost completely consist of tweets that are all original messages, others are largely made up of retweets, like the one of 2a. We consider this property in the percentage of retweets:

$$\%\ retweets = \frac{total\ count\ retweets}{total\ count\ tweets} \times 100$$

4.3 Propagation depending on Video Content

Next we examine how content and propagation correlate. If there are patterns of propagation that indicate that different types of video spread differently over a network, this shows the need to consider these differences in content for the prediction of future propagation.

Cha et al. show in [3] that videos propagate differently across the blogosphere, depending on their topic. They indicate two main categories of videos according to the way they spread. The first includes topical subjects such as news, political commentary and opinion, that spread fast and then quickly disappear. The second includes non-topical content, such as music and entertainment, that is often old but gets rediscovered and slowly gains attention every now and then. Would the categorization of propagation patterns in [3] also hold for content sharing on Twitter? If it does, it makes sense to incorporate the features in the description model for video content we presented in section 4.1 as features to build a prediction model on.

For a five-week period, we analyze differences in propagation over different categories of popular videos to inspect if we can form a similar categorization for video propagations on Twitter. We divide videos roughly according to the distinction in [3]. Topical content includes videos from more the more timely relevant categories 'Sport', 'Gaming', 'Education', 'Cars and Vehicles', 'Film and Animation', 'People and Blogs' and 'Non-profits and Activism'. Non-topical content includes the video categories 'Comedy', 'Music', 'Entertainment and 'Pets and Animals'.

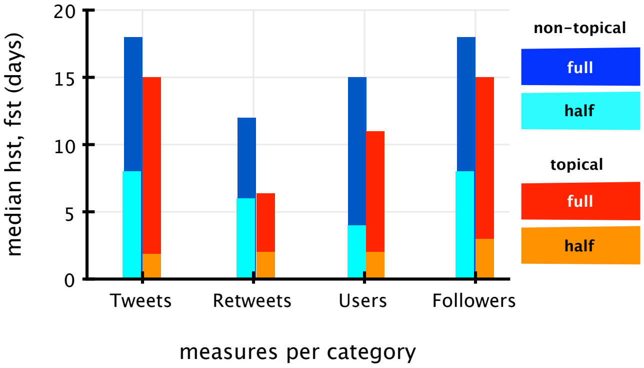
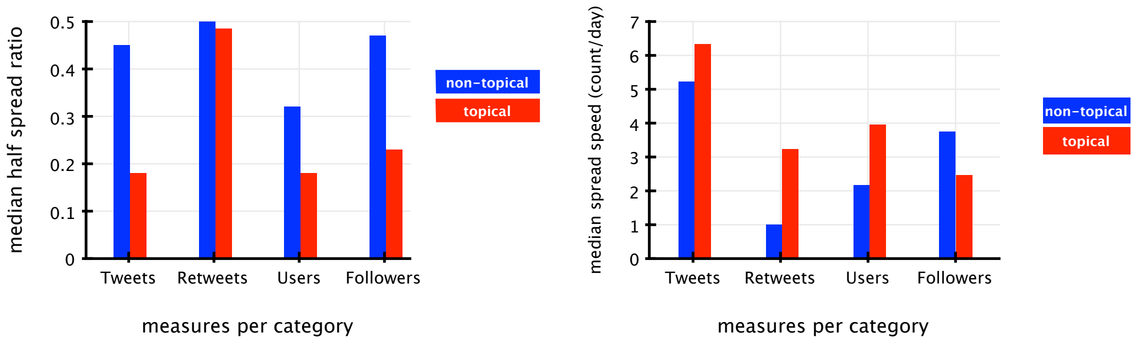


Figure 3: half (hst) and full-spreading time (fst) for topical and non-topical video categories

Figure 3 shows how both groups of videos differ according to their spreading time. Medians for both *half-spreading time* and *full-spreading time* rank consistently lower for topical content over all four measures. Topical content thus spreads over shorter periods in time and also reaches the first half of its propagation in fewer days. This last issue becomes even more clear when we look how *half-spreading time* and *full-spreading time* are related. Figure 4a presents *half spread ratios* for our four different measures. The low rates for tweets, users and followers of topical content indicate a non-uniform propagation with most activity early in the propagation period. Non-topical content shows consistently higher values, approaching 0.5 for tweets, retweets and followers. Non-topical propagations thus show a more persistent development. Interestingly, *half spread ratios* for topical retweets deviate from the trend seen in other measures, which means topical videos are retweeted in the first half of their propagation approximately as much as in the second half. Also spreading speed corroborates the categorization by means of topicality.



(a) half spread ratio

(b) spreading speed

Figure 4: measures for topical and non-topical video categories

Figure 4b shows that numbers of tweets, retweets and users increase more rapidly for topical content. Only the number of followers (plotted in *thousands/day* for readability) increases with higher speed for non-topical videos. This may be explained by the way connections are established in the Twitter network. It might be the case that certain users mention more non-topical content than others. If these users also tend to have more followers, this would explain a relatively lower spread speed of non-topical tweets combined with a higher speed for followers.

Based on the measures included in this section we conclude that the categorization by [3] of content sharing patterns by means of topicality holds for the spread of video content on Twitter. Compared to non-topical content, topical videos spread faster, over shorter periods of time and propagation tends to be more active during the start of propagations. These findings show that propagation structure is dependent on video content and therefore it makes sense to base prediction of future propagation of video in part on content features. Moreover, this shows that the features we use are well-equipped to capture important differences in propagation structure. In the next section we describe in detail how we use the two sided description model for videos to build a prediction model for video propagation.

4.4 Prediction of Future Propagation

In this section we elaborate our ways of predicting video propagation. We explain how we base estimations of the increase in propagation activity on information about video content as well as information about the propagation so far. By looking at the development of propagations in our data set, we extract examples of initial periods and their associated values of increase for a future period. We explain how we use these examples to train different prediction models by using supervised learning methods.

Video content may hint at the way a video spreads, but we do not expect video features by themselves to be expressive enough to base our prediction upon them solely. The development of a propagation over time depends largely on the activity that has taken place in the social network so far [22]. Propagation in future periods, should thus always be seen in context of its development so far. Our task of predicting future propagation of a given video, more specifically means to estimate how the four quantities tweets, retweets, users and followers will further increase based on their development so far combined with the features of the video. We measure their increase as the percentage of growth from the value after an initial period to the value after some extra time has passed.

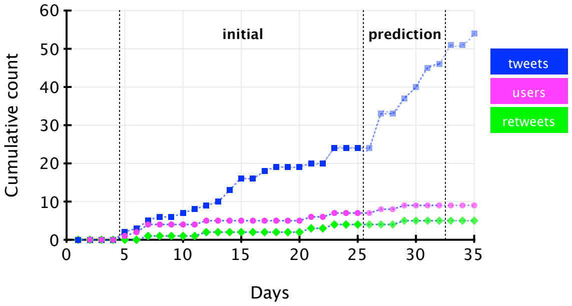


Figure 5: Illustrating the task of propagation prediction

An example propagation is plotted in figure 5 to illustrate the prediction of future propagation of a given video. Again we have excluded numbers of followers for sake of readability. In this example we take into consideration a twenty-one-day initial period, in order to predict the increase of tweets, retweets, users and followers for a successive seven-day period. By coupling the extracted propagation features for the initial twenty-one-day period to the features of the video we have a collection of input attributes that have a corresponding set of measures for the future period. Note that for a single video's propagation, like the one in figure 5 we can consider multiple initial and prediction periods of the same size. By shifting both periods in the example to the right, three more paired initial-prediction examples can be formed.

The characteristic of input attributes with corresponding values to predict, makes this task suited for application of supervised machine learning techniques. Considering examples of propagation from our data we want to learn a function that maps input attributes of video and propagation to output values of future propagation. We do so by using supervised function approximation to find a model that best fits a training set of such input output instances. We do not expect the attributes in our description model to produce good estimates for future propagation in terms of any simple linear combination. For this reason, standard methods like for example linear regression will be, although mathematically traceable, unsatisfactory to our goal. Instead, we use a multilayer perceptron network, implemented in the data mining toolkit WEKA [9], as it is able to approximate non-linear functions of input attributes, to produce continuous-valued predictions.

The task of predicting propagation one week into the future, might have different dependencies on attributes than predicting over a period of three weeks. We want to differentiate between different sizes of both initial and prediction periods and see the prediction of propagation using these varying sizes as slightly different tasks. Because of this, we build different models that each consider a set size of initial and prediction period.

In the next section we describe our experimental setup and present results for prediction performance using different sizes of initial and prediction periods.

5 Experiments

In this section, we evaluate our description models for video content and propagation. How does our prediction model perform over extended future periods? How does the performance depend on the size of the initial propagation period? We report on performance of our prediction models at varying sizes of initial and prediction periods. To begin by explaining how we acquire our data for evaluation.

5.1 Dataset

We use part of dataset of twitter content that is acquired by the University of Amsterdam and has previously been used for research on Twitter [17]. The set is acquired by daily crawling Twitter, thereby obtaining a uniform random sample (approximately 1%) from all tweets per day. Only tweet content is acquired, together with associated attributes like author, time of creation and message text. Tweets are stored in serialized .json objects, of which an example is included in appendix A. We focus on a 35-day period of this data, ranging from January 24 to February 27 2011. This complete set of data contains a total number of 31 million tweets.

Because our data is sampled at a very low rate of around 1%, videos that are shared only a couple of times a day have a small probability of showing up in our data. In order to analyze propagation over time, we take from our data the videos that are popular in terms of large numbers of tweets. These are videos that, either spread actively over possibly short periods of time or have consistent propagation extended over longer periods.

5.2 Popular Videos and their Associated Tweets

To find relevant video content and associated tweets, we take several processing steps, starting from the complete set of tweets. A diagram of our processing is shown in figure 6. As we are interested in tweets linking to video, a first processing step is to perform a coarse filtering to acquire the subset of tweets that contain a URL link to an external source. This initial filtering results in a collection of little under 6 million tweets, which is 18% of the complete set. This percentage is in line with findings in [1]. These tweets together contain 5118572 unique URLs. In the collection of tweets with URLs we index the present domain names and rank them by number of times they are tweeted. We aggregate domains that maintain multiple versions of their name (e.g. ‘google.com’ versus ‘goo.gl’ and ‘youtube.com’ versus

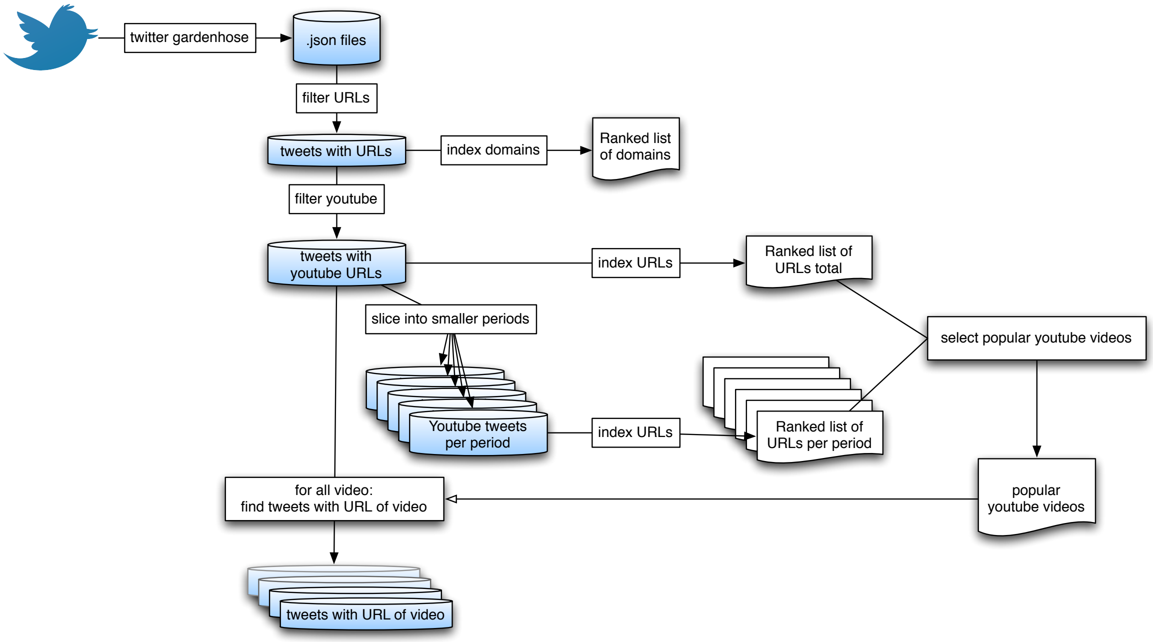


Figure 6: Framework to find popular videos and associated tweet data

'youtu.be'). This way we arrive at a top 10 of the most tweeted URL domains which is shown in table 1.

rank	**domain**	**# URLs**
1	bit.ly	1,266,182
2	tumblr.com	369,994
3	twitpic.com	227,086
4	youtube.com	209,657
5	tl.gd	156,975
6	4ms.me	155,480
7	tinyURL.com	139,327
8	google.com	125,854
9	tmi.me	123,986
10	4sq.com	120,894

Table 1: top 10 most tweeted URL domains

Some important aspects of the content sharing behavior on Twitter become clear while looking at the top 10 domains. First, URL shortener services are popularly used. At rank 1, 5 and 7 we see three of them taking up a large part of the top 10. Second, we see users sharing lots of (visual) content. Sites like twitpic.com, thumblr.com and youtube.com are examples where content is shared and externally hosted. Interestingly, Youtube shows to be the only video hosting domain that receives massive numbers of tweets. Other popular domains only barely end up in the top 100 domains. This last aspect is important to us as it shows that, when talking about online video shared on Twitter, we are mainly talking about videos uploaded to Youtube. For this reason we focus our research on Youtube videos and ignore video content that is hosted on other domains.

Focussing solely on Youtube videos, we continue to filter our data for tweets that contain youtube URLs. Here, the massive use of URLshortners might become a problem. As these shortener services shorten any kind of URL, without hinting at the original in the shortened version, information is lost about wether the URL links to a youtube video or not. The problem is partly solved by Twitter's own shortener service (t.co/...) and the way it structures tweet data into .json files. These files contain the attribute `"expanded_URL"` to indicate whether an longer version is known, in case a shortened URL is used in a tweet. However, as seen in table 1, many other widely popular services are in use. We make the assumption that shortened URLs are used with the same likelihood for different kinds of videos. By making this assumption it becomes possible to ignore shortened URLs and see the caused loss of videos for our data set as an effect of a lower sampling rate.

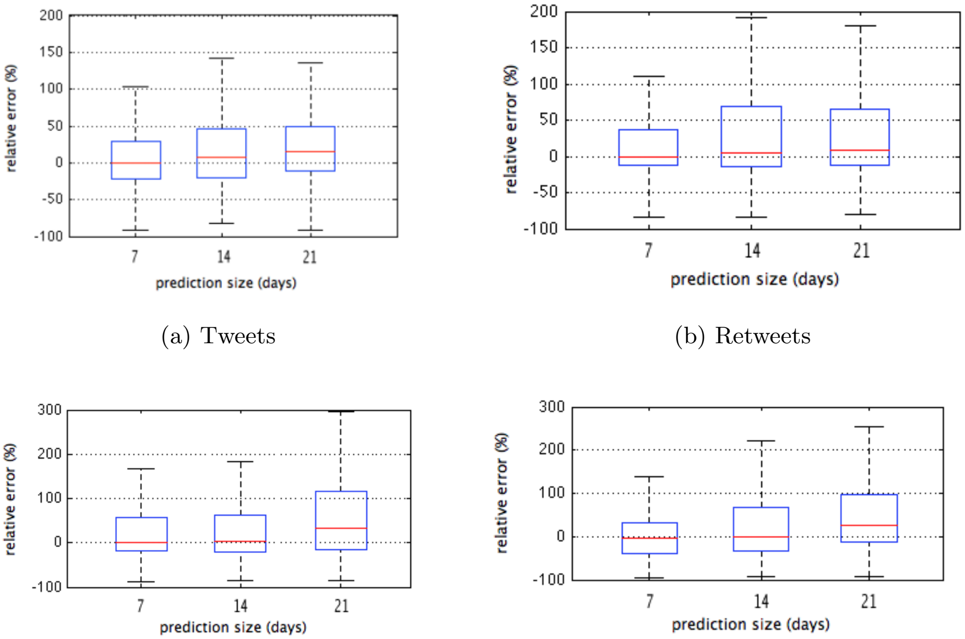
In the set of tweets containing links to videos, we index the number a video is referenced in a tweet and produce a ranked list all tweeted videos. Highly ranked videos in this list constitute the most popular videos in our complete set, but because we aggregate all videos over an entire 35-day period, videos that have a long retention time easily climb the list. In [14] Kwak et al. have shown that nearly 80% of active periods of people continuously tweeting about a topic without a 24-hour interruption, last for a week or shorter. Although their results also show that topics may have multiple active periods, this means that topics, or in our case videos with a short active period will not show up to be as popular next to topics with a long active period if measured over a longer period of time. As we aim to analyze different diffusion patterns, both of short and long time spans we extract rank lists based on video popularity for shorter time periods as well. We use five seven-day subsets of our data to find extra videos that are popular in these periods.

Finally from both the total index as well as the five indexed lists for the smaller periods we select a collection of video URLs to use in our analysis. From the total list we take the top 100 most popularly tweeted links and from the other five periods, we include the twenty most popular video of that week. In merging the selection of 200 videos, it turns out that only 8 videos from the week-long periods were not yet present in the total top 100. Upon processing these links, HTML sources of 16 videos are unavailable because their pages have been removed. This brings the size of our final set of videos to 92.

5.3 Evaluation

For all videos in our selection we find the collection of tweets that contain their URLs. These collections capture the complete propagation within our 35-day period of each of the selected videos. Together with the videos' HTML sources, these propagations are the starting point for the creation of paired initial-prediction instances based on our two-sided description model. Out of each complete propagation we form multiple subsets, that vary in size of initial and prediction periods. We use multilayer perceptron function approximation to construct from these sets, several models that predict values for tweets, retweets, users and followers.

The multilayer perceptron network uses the backpropagation algorithm to form a non-linear function of the input features that best fits the training data, and outputs a model that predicts continuous valued predictions. Although training instances will always have a minimum actual value of 100 (in case there is no increase, i.e. cumulative number stays the same), the functions estimated by the multilayer perceptron algorithm can also have lower values in its range. As performance is always improved by these values being increased, we adjust all predicted values lower than 100 to the minimum possible value of 100.



(a) Tweets

(b) Retweets

(c) Users

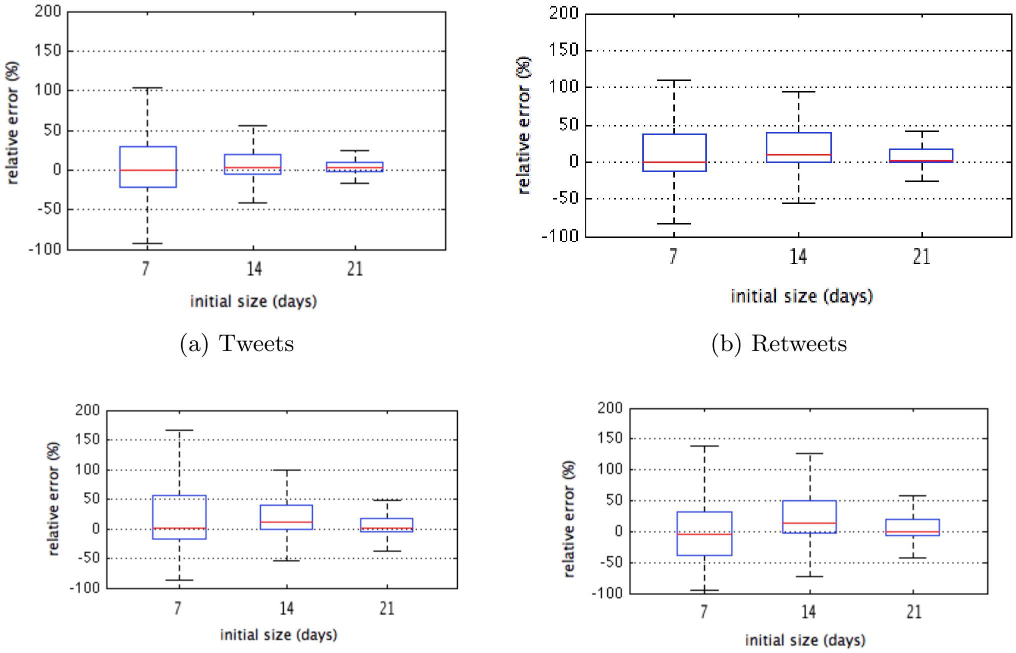
(d) Followers

Figure 7: Performance for variable sizes of prediction period

We evaluate the performance of our models by a 10-folded cross-validation. First, we measure the influence of the size of the prediction period. In other words, what is the effect of predicting further into the future? We use a steady value of seven days for the initial period on which we base our predictions for different sizes of prediction periods. Figure 7 shows how the relative error of the predictions differs for predictions of seven, fourteen and twenty-one days. We show boxplots to include information about both accuracy and variance of the predictions. The red line in the middle of the center box marks the median value of the relative error, which is more robust against outliers than the mean value. The blue box marks the area of errors between the 25 and 75 percentile, this is the area of the 50 % most accurate predictions. The two whiskers at top and bottom indicate minimum and maximum values, not considering outliers.

Overall, we can distinguish two consequences of longer prediction periods. Prediction results for the twenty-one-day periods generally have higher error values and are also less precise, as indicated by a broader variance in the error values.

Next, we examine the influence of the size of the initial period. If we base our predictions on information from longer periods of video propagations, what effect will this have for a model's performance? We predict increase of tweets, retweets, users and followers for seven-day periods, now based the predictions on a varying size of initial periods. Figure 7 shows how error values are spread for the four different aspects.



(a) Tweets

(b) Retweets

(c) Users

(d) Followers

Figure 8: Performance for variable sizes of initial period

Again, two effects can be distinguished, one more clearly visible than the other. In the figures we see a clear contraction of the boxplots as we move from shorter initial sizes to longer ones. This indicates more precision reached by predictions that are based on longer initial periods. Especially for measures of tweets and users, the smaller variance is striking. Although not immediately clear from the median values, predictions based on longer initial periods are consistently more accurate than those based on shorter ones. To make this clear, we plot *mean absolute relative error* values in figure 9.

Where the increase in prediction period only led to a slight decrease of performance in the experiments above, changing initial periods to span a longer time has a more notable effect. This may be explained by the fact that the predictions of figure 7 are all based on seven-day periods, that turn out not to have such great predictive power compared to longer periods.

The plots shown so far give insight in the performance of how our models perform, mainly relatively per category. To compare the predictability of these different attributes to each other we take the best performing model for each of the four attributes and include a more detailed *cumulative distribution function* for absolute relative errors in figure 10. These best performing models all use initial twenty-one-day period to predict propagation in future seven-period.

Best prediction results are achieved for tweets. Nearly 90% of the predictions for tweet propagation are within an relative error margin of 20%. Next to this, almost 70% of the predictions is even within a margin of 10% relative error. Interestingly, tweet predictions do

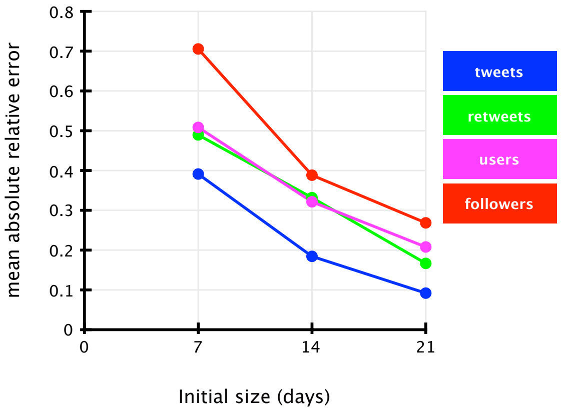


Figure 9: Mean absolute relative error for variable sizes of initial period

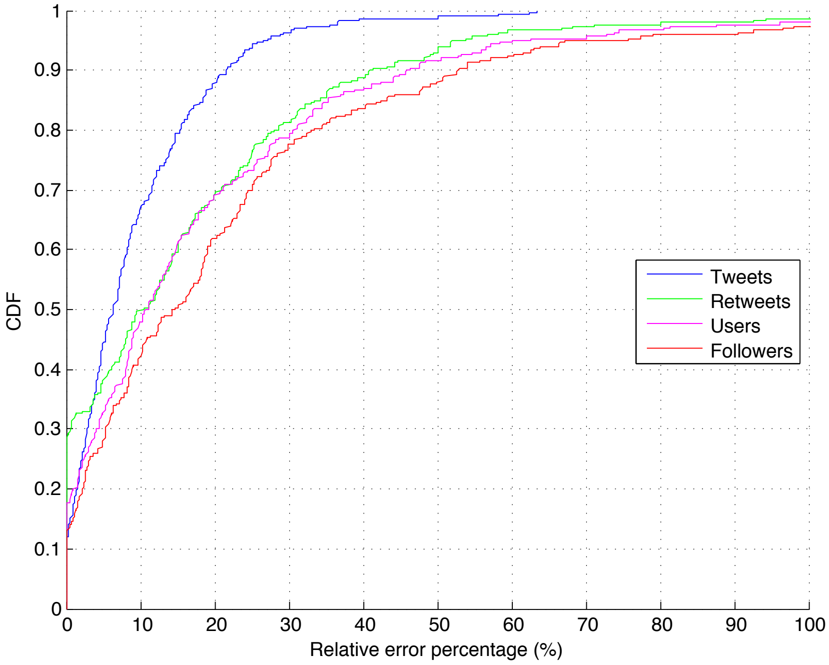


Figure 10: Cumulative distribution plot for best performing models

not show serious outliers, but stay all within the range of 70% relative error. This is unlike the other categories, such as followers where predictions show errors up to 1400% relative to the actual value.

6 Conclusion

We have presented a generally applicable, two-sided model to describe video content and its propagation on a social network. We have investigated correlations between the content side and the propagation side of our model. General patterns of content-dependent propagation, that have earlier been found to be present on blogs, are applicable to video sharing on Twitter. This content-dependency of propagation, shows the need to include video features when building a coherent prediction model.

We have used the supervised non-linear function approximation methods to build prediction models that are able to predict increase of tweets, retweets, users and followers. In the evaluation of our prediction models we have shown the predictive capabilities of our features and indicated the influence of different sizes for initial and prediction period. These measures determine how much time backwards and forwards a model considers in predicting future propagation. We have indicated the difficulties of predicting propagation further in the future and have shown strong benefits of considering an extended initial period.

The fact that we use sampled data has prohibit us from analyzing the Twitter network on a node to node level that considers Twitter's social graph of inter-personal connections. Although our methods prove quite predictive, there is no doubt that our propagation model can be meaningfully augmented by more local attributes that capture propagation features to the level of individual cascades. Also real world knowledge about external events that influence a propagation, such as a feature on another website of network, might be a valuable enrichment.

References

[1] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *hicss*, pages 1–10. IEEE Computer Society, 1899.

[2] M. Cha, A. Mislove, and K.P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730. ACM, 2009.

[3] M. Cha, J.A.N. Pérez, and H. Haddadi. Flash floods and ripples: The spread of media content through the blogosphere. In *ICWSM 2009: Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*, 2009.

[4] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1029–1038, New York, NY, USA, 2010. ACM.

[5] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs.

In *Proceedings of the 3rd conference on Online social networks*, WOSN'10, pages 3–3, Berkeley, CA, USA, 2010. USENIX Association.

[6] D.R. Gibson. Concurrency and commitment: Network scheduling and its consequences for diffusion. *The Journal of mathematical sociology*, 29(4):295–323, 2005.

[7] M.S. Granovetter. The strength of weak ties. *The American journal of sociology*, 78(6):1360–1380, 1973.

[8] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[10] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[11] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[12] S. Kinsella, A. Passant, and J. Breslin. Ten years of hyperlinks in online conversations. 2010.

[13] Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 435–443, New York, NY, USA, 2008. ACM.

[14] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[15] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. *Advances in Knowledge Discovery and Data Mining*, pages 380–389, 2006.

[16] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[17] K. Massoudi, E. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *ECIR 2011: 33rd European Conference on Information Retrieval*, pages 362–367, Dublin, 2011. Springer, Springer.

[18] M.E.J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002.

[19] E.M. Rogers. Diffusion of innovations. 1962.

[20] E. Sadikov and M.M.M. Martinez. Information propagation on twitter.

[21] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 55–64. ACM, 2011.

[22] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. *Proc. ICWSM*, 2010.

[23] Youtube. Youtube press timeline, 2011.

A Example Tweet Entry

```
{
 "user":{
    "follow_request_sent": null,
    "profile_use_background_image": true,
    "id": 15998753,
    "verified": false,
    "profile_sidebar_fill_color": "FFA408",
    "profile_text_color": "333333",
    "followers_count": 739,
    "profile_sidebar_border_color": "FFA408",
    "id_str": "15998753",
    "profile_background_color": "FFF8B2",
    "listed_count": 56,
    "utc_offset": -28800,
    "statuses_count": 13697,
    "description": "Life is what happens to you while you’re busy making other plans",
    "friends_count": 630,
    "location": "Nye County, NV. U.S.A.",
    "profile_link_color": "8B8B8B",
    "profile_image_url":
        "http://a2.twimg.com/profile_images/1195741119/Copy_of_Copy_of_Copy_of_KittyTeeth_normal.GIF",
    "notifications": null,
    "show_all_inline_media": false,
    "geo_enabled": true,
    "profile_background_image_url":
        "http://a3.twimg.com/profile_background_images/184526901/Hands_Off_Wikileaks.jpg",
    "screen_name": "TurboKitty",
    "lang": "en",
    "following": null,
    "profile_background_tile": true,
    "favourites_count": 27,
    "name": "  TurboKitty",
    "url": "http://www.grassrootsinternetradio.com/",
    "created_at": "Tue Aug 26 16:32:00 +0000 2008",
    "contributors_enabled": false,
    "time_zone": "Pacific Time (US & Canada)",
    "protected": false,
    "is_translator": false
 },
 "favorited": false,
 "contributors": null,
 "truncated": false,
 "text":
   "Like remorina, the tears are flowing from my eyes and I can’t tell why except...
     (YouTube http://youtu.be/ThvBJMzmSZI?a)",
 "created_at": "Fri Jan 28 17:12:56 +0000 2011",
 "retweeted": false,
 "in_reply_to_status_id_str": null,
 "coordinates": null,
 "in_reply_to_user_id_str": null,
 "entities": {
     "user_mentions": [],
     "hashtags": [],
     "urls": [
         {"url": "http://youtu.be/ThvBJMzmSZI?a", "indices": [90, 119], "expanded_url": null}
     ]
 },
 "in_reply_to_status_id": null,
 "id_str": "31037019148787713",
 "place": null,
 "in_reply_to_user_id": null,
 "in_reply_to_screen_name": null,
 "retweet_count": 0,
 "geo": null,
 "id": 31037019148787713,
 "source":
   "<a href=\"http://www.google.com/support/youtube/bin/answer.py?hl=en&answer=164577\"
     rel=\"nofollow\">Google</a>"
}
```