Will this Celebrity Tweet Go Viral? An Investigation of Retweets

L.Gandy¹, L. Hemphill²
Central Michigan University, Computer Science Department
Illinois Institute of Technology, Lewis College of Human Sciences
gandy1l@cmich.edu, libby.hemphill@iit.edu

Abstract

Celebrities have some of the largest followings on Twitter, and their fans increasingly use Twitter as a way to engage and extend celebrities’ reach. We examine retweeting, specifically, and analyze the features of celebrity tweets that predict whether it will be widely retweeted by followers. We used both multiple regression and classification approaches and found that the number of followers of a celebrity tweeter is the highest predictor of whether a tweet will be retweeted at high volume. Multiple linear regression was unable to pinpoint exactly when a tweet becomes successful but using a Generalized Liner Models classification function, we are able to predict whether a tweet will go viral with 78% prediction accuracy compared with a 66% baseline. We explain these computational approaches to analyzing millions of retweets and discuss the implications of this retweeting activity.

Previous work has been done on which tweets are retweeted in the Twitter social networking space [17][18][19][20][21]. There have been some small scale studies examining celebrities retweeting behavior, but to our knowledge there have been no large scale studies. In fact large scale studies have not focused on celebrities, but instead have focused on behavior of non-celebrity tweeters. Social Media outlets, such as twitter, has fundamentally changed the way in which celebrities and their fans interact. In the past, celebrities spoke through public relations firms, and their message was highly regulated. With the advent of social media, celebrities now regularly address and interact with their fans in a more personal manner.

When examining celebrity tweets, our two fundamental research questions were Which tweets are retweeted more than others? and What do these retweets, if anything, have in common?. Therefore, the goal of this paper is to illuminate features of celebrity tweets that are widely retweeted in order to understand what influences whether a celebrity’s comments will reach larger audiences. We examine retweets of celebrities’ tweets for two primary reasons. First, as offline celebrities’ social media audiences (and potential to influence) grow, it is important to understand what influences whether their messages are passed on. Second, as more scholars study celebrity-fan interactions, it is useful to identify what, if anything, about celebrity twitter accounts or celebrities’ tweets influence whether users will interact in the form of retweeting.

1 Related Work

1.1 Celebrities, Fans, and Twitter

Twitter provides a unique kind of conversational space that is at once both public and personal. Users’ tweets are public, in that anyone can read them, and personal in that the message can be directed at a person or group or contain personal content. Twitter also allows one-way, unreciprocated connections enabling users to view messages from people who don’t view messages from them. Most celebrities follow far fewer accounts than they have followers, suggesting that they are taking advantage of the non-reciprocal connections Twitter allows. This kind of one-sided relationship is a hallmark of parasocial interaction [1][8], the kind of relationship a media user develops with a media figure. Scholars have recently begun to examine parasocial interaction in social media among fans and sports figures [10], celebrities [12] and musicians [2]. Celebrities are the most-mentioned and most-retweeted users on Twitter [5] and have tremendous opportunity to influence what gets talked about.

1.2 Why Tweeters Retweet

Findings from Boyd’s [3] work suggest that people retweet to amplify or spread retweets to new audiences, to inform a specific audience, to comment on someone else’s tweet, to make their presence as a listener visible, to agree with another twitter user, to validate their and other’s thoughts, as an act of friendship or loyalty, and for self gain.

Macskassy and Michelson [11] found that users are more likely to retweet content from someone they have talked to or about recently, a topic they often tweet about, or a user to whom they are similar.

Yang et al [15], find that though some properties of tweets predict whether a tweet will be retweeted extensively, the most accurate predictor of whether the tweet will be retweeted is how often the original tweeter is mentioned on twitter in general. That is, the tweeters popularity, as measured by mentions, is the greatest predictor of whether his or her tweets will be retweeted to a great extent.

Conversely, Cha, et al, [5] do not specifically look at why tweets are retweeted, but rather examine how to measure how influential a twitter user is, that is how to predict which tweets will go viral, when the only information we have is who is the originator of the tweet. They find that retweets are not driven by the number of followers a user has, so
conclude that retweets must have something to do with the content value of the tweet itself. However they find that the top influentials were recognizable public figure i.e. celebrities.

2 Methods

2.1 Past Work

There have been several different statistical and algorithmic approaches to studying retweeting.

Boyd [3], used a qualitative measure of why and when users retweet. She asked questions such as What do you think are the different reasons for why people RT something?, and What kinds of content are you most likely to retweet? using the @zephoria public twitter account. She received 99 and 73 detailed responses for each questions, respectively.

Macskassy [11], et al. first generated topic profiles of twitter users. They found these topic profiles by finding the entities in each user’s tweets. They then mapped the entities onto high-level topics by using Wikipedia’s folksonomy. Once a twitter user’s topic profile was created, their retweets are examined for homophily against their twitter profile topics.

Yang, et, al [18], examined social networks and spread of tweets throughout the social network specifically. Therefore, given a social network G and a set M of tweets and retweet behaviors in history, Yang et, al. aimed to predict if users will retweet the tweet m to their friends after viewing it, and the range of spread for a new tweet m written by user u. They then measured retweet influence by noting the number of retweets mentioning a twitter user’s name. They restricted their study to twitter user’s who had tweeted ten or more times. They focused on six million twitter users. They ranked each user as influential by ranking all users using Spearman’s rank correlation coefficient.

2.2 Our Method

2.2.1 Data Collection

To being our work, we collected a list of the top 5000 celebrities ranked by StarMETER [15] ratings, according to the Internet Movie Database (IMDb) [9] on October 13, 2013. StarMETER is the proprietary list that IMDb compiles continuously. Its ranking algorithm functions by tracking IMDb celebrity page views as well as celebrity mentions in news, among other measures. We then used an automated google search to find related twitter accounts per each celebrity name. We then used the Twitter API to collect additional twitter profile information and tweets from all user accounts which we collected fusing Google IMDb. There were more 12,050 accounts that matched names from IMDb’s list. Therefore, to disambiguate user accounts we first did simple name matching. For example Miley Ray Cyrus matches Billy Ray Cyrus’s twitter account using google, therefore a simple name match comparing Miley Ray Cyrus against Billy Ray Cyrus would eliminate the extraneous Billy Ray Cyrus account. After disambiguation of celebrity twitter accounts, we found a subset of 3665 accounts that most likely belonged to celebrities on the list.

Of course not all of the tweets posted to those accounts are actually posted by the celebrities themselves. Many celebrities use ghost tweeters, and it’s difficult to distinguish between accounts authored by the individual celebrities and those with ghost or multiple authors. Many Twitter accounts that seem like celebrity accounts are actually maintained by fans or others outside the celebrity’s personal media team. Though we did not examine accounts for markers that indicate the celebrity as primary author [12], we did remove accounts that seemed fake or fan-authored. To find fake or fan-authored accounts, we looked at celebrities who appeared to have more than one twitter account in our local database. We marked the account which had the most twitter followers as valid. After this there was a remainder of 2867 twitter accounts for celebrities. In general, we treat the Twitter account associated with the celebrity as a proxy for the individual and are interested in what happens after the tweet is posted. Whether it matters for retweeting behavior if a tweet was posted by the individual or his/her proxy is reserved for future work. In all, for valid users we collected 4,304,415 tweets.

2.2.2 Data Analysis

For each celebrity twitter account that was deemed valid via the prior section, we collected the number of followers for the celebrity account, as well as other basic information about the celebrity account, such as the profile URL, profile synopsis (if any), and profile image. In relation to celebrity information, we also collected celebrity tweets per account via the Twitter API. For each tweet we also recorded whether the tweet had a URL and photo associated with it, as well as if the tweets contained mentions. We then further analyzed each URL to distinguish if any urls contains within the tweet pertained to photos or videos. We ascertained if the tweet featured engagement by a simple search for question marks within each collected tweet. In [Table 1] below, we list features that we included in our data analysis, for categorical data we list the percentage of all collected tweets which include the category in question.

2.3 Results

2.3.1 Linear Models

We started with a straightforward linear regression model using number of followers as the only predictor of retweets (Model 1 shown below in [Table 2]). Model 2 includes dummy variables for has hashtag, has mention, has URL and has photo as predictors as well. In both models, followers are a strong predictor of retweets ($p < 0.001$). From Model 1 to Model 2, we see little improvement in the overall model ($R^2 = 0.14$ to $R^2 = 0.15$) though has hashtag, has mention, and has URL are significant ($p < 0.001$). We then calculated a retweets per follower ratio and used that measure as the dependent variable (Model 3). The overall effect sizes for the models are not meaningful (0.004 - 0.18).
2.3.2 Classification Approach

To begin classification, we divided celebrity tweets into those that were deemed viral and those who were not viral. To find viral tweets, we used the \( MAD_e \) method for outlier detection. The \( MAD_e \) method, using the median and the Median Absolute Deviation (MAD), is one of the basic robust methods for outlier method, and it is for the most part not affected by the presence of extreme values in the data set [4]. The \( MAD_e \) method is defined below in Equation 1.

\[
MAD_e = Median \pm 2MAD
\]

where

\[MAD + e = 1.483 * MA\]

for large normal data and

\[MAD = median(|x_i - median(x)|,1,2,...,n)\]

Equation 1: MAdE Outlier Detection Method

MAD is an estimator of the spread in a data, similar to the standard deviation, but has an approximately 50% breakdown point like the median. Since this approach uses two robust estimators having a high breakdown point, i.e., it is not unduly affected by extreme values even though a few observations make the distribution of the data skewed, the interval is seldom inflated, unlike the SD method [14].

The \( MAD_e \) method was used on a per- celebrity basis. That is we judged whether a tweet was viral for a single celebrity, not as a global comparison between celebrities. When using the \( MAD_e \) method we found that 76,098 of a total of 4,304,415 celebrity tweets were viral.

Because of the large imbalance between viral and not viral tweets, we use the SMOTE method for data balancing [6]. SMOTE uses both over and under sampling to balance out data sets. SMOTE’s method of oversampling the minority class and under-sampling the majority class achieves better classifier performance than varying loss ratios via Ripper or Naive Bayes class priors. After using SMOTE, the data set consists 1,217,568 tweets which are labeled as viral and 2,282,940 tweets which are labeled as not viral. We then use the Generalized Linear Model [13] for Classification. The glm model predicts whether a tweet will be viral or not with a precision of 78%, recall of 99% and a f-measure of 87%. Due to the balance of our data there would be a precision of 66% if all tweets were counted as not viral, our classification model then improves on the random model by 14%.

2.3.3 Feature Selection

To further investigate the importance of features in classification, we used the Recursive Partitioning (RP) Decision Tree [7] to perform feature selection. Recursive portioning works by recursively splitting the data and finding the decision tree which best fits the data. In the process, the recursive partitioning algorithm also performs feature selection. Recursive portioning classified data with 76% accuracy which is quite close to the predication accuracy achieved by the generalized linear model binomial classifier. In addition the RP algorithm only uses the number of followers feature to achieve this accuracy. The RP algorithm tree classifies tweets whose associated celebrity tweeater has more than 1,614,260 followers as viral and less than 1,614,260 followers as not viral.

3 Discussion

To our knowledge this is the first large scale study of celebrity retweeting behavior. Therefore we have no other studies to compare this work to. However we can compare
results to large scale studies of non celebrity retweeting behavior. Though this comparison is not one to one it will give us a starting place for discussion.

We feel that our results do not contradict the findings of Yang et al [18]. Yang finds that the most accurate predictor of a tweet being retweeted is how often the original tweeter is mentioned on twitter. We would posit that frequent mentions on twitter indicate a twitter user is a twitter celebrity. We would also posit that our finding that a twitter user who has a large number of users also marks a twitter user as a twitter celebrity. Therefore both Yang, et al, and our findings indicate that whether a user is or isn't a twitter celebrity indicate whether their tweets will be retweeted. In our case, however, our twitter users are also celebrities outside the twitter sphere.

Our findings also do not disagree with those of Macskassy and Michelson [11], who found that users are more likely to retweet content from someone they have talked to or about recently, or a user to whom they are similar. It is not a huge assumption that a twitter user would talk about their favorite celebrities frequently. Therefore we are not surprised that tweets from celebrities with large number of followers are retweeted extensively. Also, Boyd et al [2] speak about the fact that celebrities often tweet to gain a sense of affiliation or similarity between themselves and their audience, it is also not a stretch to then imagine that twitter users would feel a similarity or kinship to celebrities that they admire and thus retweet those celebrities’ tweets.

We believe that the difference in our finding’s and those of Cha’s [5] exist primarily because of the different methods we used to explore our datasets. First, we only looked at retweets that were retweeted much more than other tweets. However Cha, et al examine all user retweets. Cha also focuses on the how influential a user is, where as we do not so much fixate on the user but the tweets themselves. This would explain why our findings most closely mirror those of Yang, et al, and Macksassey et al, who both ask Which tweets are retweeted whether than Which twitter users are influential?

4 Future Work

In our future work we plan to either hand label or automatically categorize tweets into types. We would like to examine whether the semantic content of a celebrity weet affects its propagation via retweets. In addition we plan to further categorize celebrities by level of celebrity via their ranking on IMDB. We would like to gauge whether the level or celebrity outside of social media affects retweets inside the twitter verse.

In addition, we would like to collect an even large range of tweets, whereas now we collected just over 4 million tweets, we would like to collect on the order of 1 billion tweets. This will aid us in our future work exploring types of tweets and will also further validate our current work on a larger dataset.

5 Conclusion

In our study of celebrity retweeting, we began with a list of 5000 celebrities taken from IMDB. We then performed data mining and ended up with 2867 valid celebrity twitter accounts In all, for all valid celebrity tweeters we collected 4,304,415 tweets. We divided tweets into those that were not viral and those which were viral by using the MADE outlier detection method. We then used regression, classification and feature selection to explore what about the tweets or tweeters would guarantee a tweet going viral. We found overall that the number of followers that a celebrity tweeter has is the most solid predictor of whether a tweet will go viral.

References


