Predicting ReTweet Count Using Visual Cues

Ethem F. Can and Hüseyin Oktay and R. Manmatha
School of Computer Science
University of Massachusetts Amherst
140 Governors Dr., Amherst, MA, USA
{efcan,hoktay, manmatha}@cs.umass.edu

ABSTRACT

Social media platforms allow rapid information diffusion, and serve as a source of information to many of the users. Particularly, in Twitter information provided by tweets diffuses over the users through *retweets*. Hence, being able to predict the retweet count of a given tweet is important for understanding and controlling information diffusion on Twitter. Since the length of a tweet is limited to 140 characters, extracting relevant features to predict the retweet count is a challenging task. However, visual features of images linked in tweets may provide predictive features. In this study, we focus on predicting the expected retweet count of a tweet by using visual cues of an image linked in that tweet in addition to content and structure-based features.

Categories and Subject Descriptors

H.2.8 [**Database and Management**]: Data Mining; J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords

Twitter, Retweet prediction, Social Media, Visual Cues

1. INTRODUCTION

Social media platforms enable users to not only share information with their peers but also discover new information. Such information reaching to large populations has the potential to influence public opinion [1], market share for new products and the adoption of innovations. To maximize the reach of a certain piece of information or product to large populations, researchers and marketers have become interested in using various different strategies on social media such as seeding information with 'influentials' [4]. One way to maximize such efforts could be by predicting the diffusion of information on social media to quantify the expected exposure of different strategies. In particular, for example, the basic mechanism for the diffusion of such information over Twitter is through retweeting. Users can choose to retweet a particular tweet to share the content with their followers. Hence, understanding the mechanisms of information diffusion and predicting the retweet count beforehand are important tasks to quantify the expected reach of a particular piece of information. The prediction of the retweet count is a challenging task, partly because Twitter is a complex domain with many users interacting with each other by generating massive content on a daily basis.

A number of qualitative and quantitative analysis have been performed towards understanding information diffusion and influence on Twitter. Kwak et al. [5] qualitatively compared different definitions of influence and mainly showed that influence ranking depends on the definition. Boyd et al. [2] formulated retweeting as a conversation practice to qualitatively understand the underlying mechanism for retweeting. Yang and Counts [10] compared diffusion structures on Twitter to diffusion structures on web logs to identify qualitative differences and similarities. Quantitatively, Suh et al. [8] used generalized linear models and identified content-based and structure-based features that are significantly correlated with retweet count such as having a link in a tweet and the number of followers of the user who posted the tweet. Yang and Counts [11] used survival analysis to predict the scale, speed and range of tweets by focusing on the diffusion of certain topics on Twitter network. Bakshy et al. [1] focused on predicting the influence of individuals on Twitter by considering the cascades of shared urls. In their predictive model they used manually labeled content-based features along with structural features and the past influence of a user to predict influence from a user perspective. Other problem formulations from a user's perspective include the following papers which predict the retweet count of a user. Zaman et al. [13] used a collaborative filtering model utilizing content- and structure-based features. Yang et al. [12] used a factor graph model. Uysal and Croft [9], also predicted the likelihood of a retweet by a user on a given tweet with a logistic regression model and used the information to rank the tweets according to their interest to a user.

One common constraint reflected in such studies is that Twitter is a limited domain where a tweet's content cannot be longer than 140 characters making it hard to come up with predictive content-based features. From the perspective of users, such a constraint is about the level of information that a tweet can contain. Hence, users use url links to websites, news articles, pictures or even short videos to at least refer to more content in their tweet. Users make use of url shorteners while they are tweeting so that they do not use too many characters (e.g., bit.ly), or they make use of photo hosting platforms combined with url shorteners to

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
Copyright 2013 ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2507824.

include images and short videos in their tweets. As users include more information in their tweets by including pictures, we can make use of such additional information contained in the shared images in predictive models to estimate the retweet count for a given tweet.

In this paper, we focus on tweets that contain links to images shared through twitpic.com, and extract correlated low-level and high-level image features in those tweets. We, then, use these correlated visual cues to increase the accuracy of our predictive model for retweet count. Our low level features are global color histograms and GIST features [7] while our high level features use the set of responses to specific object detectors [6]. We, then, compare three different regression techniques (linear regression, support vector machines and random forests) over these features as predictive models to estimate the retweet count of a given tweet.

Our contributions in this paper are three-fold. First, we introduce followers-to-friends ratio as another feature highly-correlated with the retweet count, and show that random forest regression model including this new feature along with additional content- and structure-based features suggested in the literature gives a statistically significant improvement in performance. This new combination of features serves as our baseline set of features in our experiments. Secondly, we augment our baseline features with image features either color, GIST or high-level features and show that the *Pearson's correlation* of some of the visual features is highly-correlated with the retweet count. Finally, we show that using such image features in our regression models we can substantially increase the accuracy of predicting the retweet count for a given tweet. By increasing the accuracy of retweet count prediction, we enable a better analysis of information diffuse through tweets exploiting visual information linked in the tweets as well.

2. FEATURES

In our predictive models we focus on three types of features; content-, structure-, and image-based features. They are summarized in Table 1.

Table 1: Summary of features

	Name	Value
Content-Based	hasHashTab	binary
Structure-Based	followersCount	numeric
	friendsCount	numeric
	followersFriendsRatio	numeric
	age	numeric
	statusCount	numeric
	favoritesCount	numeric
	listedCount	numeric
Image-Based	colorHistogram $cf_0, ..., cf_{511}$	numeric
	GIST $gf_0, ..., gf_{511}$	numeric
	objectDetectors $obf_0, ..obf_{176}$	numeric

Content-Based Features: *hasHashTag* is a binary feature that represents the presence or absence of a *hashtag* in a tweet. Our target variable of interest is the *retweetCount* of a given tweet that is content-based as well. The *retweetCount* has a power law distribution. To avoid the bias from one data instance with a high number of *retweetCount*, we transformed the data to log scale, and use the log value in our predictive models.

Structure-Based Features: *followersCount* is the total number of followers of the owner of a tweet. When a user sends a tweet, all the followers of that particular tweet get that particular tweet in their news feed. Hence the more the followers a user has, the wider the exposure of tweets of that user. *friendsCount* is the number of friends that the owner of a tweet (i.e. the number of people that the owner follows). A user gets all the tweets from her friends. Hence, the more friends a user has, the more tweets a user gets. To again avoid biasing the scores towards users with a high number of followers in our models, we also transform the values of these features by using the log of the feature values. *followersFriendsRatio* is the ratio of the number of followers to the number of friends for the owner of a tweet. Such a ratio can be considered as a measure of the activity of a particular user. *age* is the total number of days the user has been active until a tweet is posted. Another structure-based feature is *statusCount* which is the total number of tweets of a user until the particular tweet is created (i.e., the total number of tweets tweeted by a user). We transform the feature values into log scale. *favoritesCount* and *listedCount* are relatively recent features that have become available on Twitter data. *favoritesCount* is the number of tweets a user marked as favorite and, it has also a power law distribution. *listedCount* is the number of public lists that a user is a member of and it also follows a power law distribution. Hence, we transform both of the feature values into log scale.

Image-Based Features: We consider three image based features; two low-level features and an object based feature [6]. First, *color histograms* provide information about the distribution of color intensities. We extract color histograms as our image features in the following way. For each channel of a color image, i.e. red, green, and blue, we uniformly quantize the range of color intensities into 8 distinct values. Second, *GIST* descriptors provide a set of perceptual dimensions such as naturalness which represents the major spatial structure of an image. We compute a GIST descriptor for each image in the following way. The image is subdivided into 16 regions (i.e., 4 by 4 grids) and there are 8 orientations at each scale. The number of scales is set to 4 in our case. The third feature is the set of responses of individual object detectors $\{obf_0, .. ., obf_{176}\}$ which are used in our predictive models as well [6]. In our object bank, there are object detectors for 177 different objects such as dog, car, and table. Note that for each object we have a number output by the corresponding detector. Having computed and max-pooled the responses, for each image we have a 177-dimensional histogram.

3. EXPERIMENTAL SETTINGS

In our experiments, we compare two different cases: *baseline* and the baseline augmented with image-features. As a baseline we use content- and structure-based features, which are explained in Section 2. Such features include almost all the features reported in the literature as being highly correlated with the retweet count. Different from the literature, we also include *followersFriendRatio* in our baseline feature set which is shown to be highly correlated with the retweet count.

We crawled Twitter data using the Twitter sampling API (https://dev.twitter.com/docs/streaming-api/concepts) for two weeks. The Twitter sampling API samples roughly 1% of all tweets from the Twitter firehose (i.e., full feed of public tweets). A random sample preserves the distribution of tweets. We sampled the tweets that shared a picture using twitpic.com which is the most commonly used picture sharing platform on Twitter [8]. Among 140K tweets in our sample with a link to pictures, we were able to extract visual features for more than 100K of those tweets (some of the links did not contain images hence we eliminated those tweets from our sample). Among 100K tweets, we create a training set by randomly sampling 10K tweets and we sample a different set of 10K tweets to create the test set. We train our model on the training set, and we evaluate the performance of our model on the test set. We replicate this sampling process 100 times and we evaluate our results using the *root square mean error* (RMSE) metric. For this metric, the lower the value, the better the performance. All the results in this paper are based on this evaluation metric.

In our experiments, we make use of the R statistical software package. For SVM we use kernlab version 0.9-14 package that is already implemented in R. For random forest regression we use the randomForest package version 4.6-6 which is also implemented in R. We use default parameter settings.

4. RESULTS AND DISCUSSION

We formulate the problem of predicting the retweet count of a given tweet as a regression problem. Given a set of features F, and a target variable T, we experiment with different regression methods to predict a target variable. Our target variable is the *logarithm* of the retweet count for a given tweet. Our feature set F contains all the features explained in Section 2. In our experiments, we focus on three different types of regression; linear, SVM with a Gaussian kernel, and random forest [3].

In Figure 1, we compare the baseline and our approach for different regression methods (i.e., linear regression, SVM, random forest respectively). At a high-level, independent of the regression method, visual features extracted from object responses improve the prediction accuracy of the retweet count for a given tweet. Low level features also improve the prediction accuracy for some classifiers. We observe the most improvement with the random forest. Experiments with baseline features provide a root mean square error (RMSE) score of 1.743, 1.722, and 1.553 in *log* scale and 5.713, 5.599, and 4.725 in linear scale) for linear, SVM, and random forest regressions respectively (note that a lower RMSE is better). With our approach using object based features the RMSE scores down to 1.703, 1.559, and 1.297 (5.489, 4.753, and 3.659 in linear scale). The results with our approach are statistically significantly different from the results of the baseline with three different regression techniques.

To reduce the variance, we slightly modify our approach to discard features that are not correlated enough with retweet count (i.e., below a threshold value=0.05). We labeled this case as *cutoff* in Figure 1. For object-based features, from a total of 177 features, by thresholding we use only 60 features that are correlated enough with retweet count. Our conclusion is that we observe slight improvements by thresholding features with linear regression, whereas the performance is not statistically significantly different for SVM and random forest models. To understand which features help we compute the correlations of the baseline features and the ObjectBank type features with the retweet count and sort the values in terms of significance. The most correlated feature is *followersCount*. This outcome is not surprising since the more followers, the more users a tweet reaches, and hence the higher the retweet possiblity. Surprisingly there are 16 image-based features in the list of top 20 most correlated features and many of the structure-based and content-based features don't appear in the list of top 20. Having some of the object-based features in the top ranks explains why visual cues improve the accuracy of prediction models. The only feature which is negatively correlated is the *age* feature.

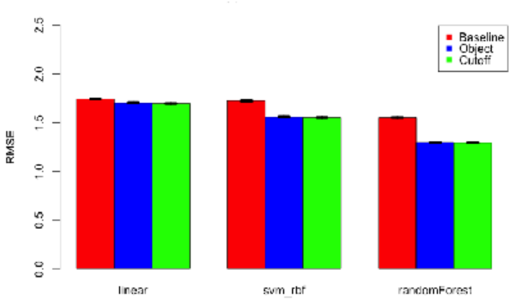

Figure 1: Prediction results of the models with the baseline features compared with the models with baseline features and object-based features (also cutoff experiments) on three different regression techniques. Lower is better.

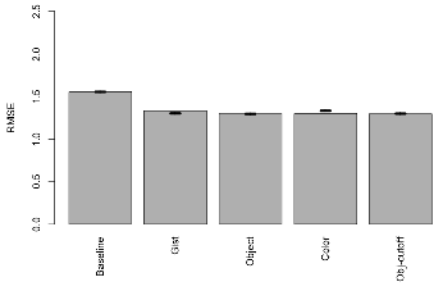

Figure 2: Prediction result of the models with the baseline features compared with the models with the baseline features and different image-based features when random forest regression is used.

In Figure 2, we compare the results of baseline with the low-level and high-level image image-based features with random forest regression. The results show that using a random forest, low-level image-based features also improve the prediction accuracy of retweet counts. The prediction model using low-level color histogram features perform as well as high-level object features. The prediction model with low-level GIST features outperforms the accuracy of the model using just baseline feature. However, GIST features are not as good as color histograms or object-based features. Note that all image-based features are combined with the baseline features.

When we consider the results of random forest regression, the best models include baseline features combined with color-based features or baseline features combined with object-based features. The model which combines baseline features with GIST descriptors is slightly worse and the model with baseline features is substantially worse and is the worst performer. Here, we claim that visual cues independent of whether they are low-level or high-level features improve the predictive power of our model using random forests. However, when we consider different regression techniques the results are not the same. The models with color-based features + baseline or GIST descriptors + baseline are not better than the model with baseline features alone when using linear regression or SVM. Models with object-based features always performs better than the model with baseline features independent of the regression method we use. Object-based features provide more robust results than other low-level image features independent of the regression technique.

With random forest regression, we get an RMSE value of 1.297 on a log scale, by improving the accuracy over baseline 16%. Since the difference on a log-scale, is division in linear scale, our RMSE value for random forest corresponds to 3.66 times, and the baseline corresponds to 5.49 times in linear scale. For example, if the actual retweet count of a tweet is 100, baseline predicts a value roughly between 18 and 549, whereas random forest model predicts between 27 and 366. Although such prediction results may not seem perfect enough, the results significantly improve by including visual features from images and that is the main purpose of this paper. We conjecture that the more accurate the features extracted from images, the more accurate the performance of prediction of the retweet count will be.

Currently, we consider responses to object detectors as one of our image-based features, and we observe that they are more correlated with retweet count than some of the structure based features. Responses of object detectors such as flower and ferris-wheel are positively correlated with retweet count.On the other hand, responses of object features such as building and blind are negatively correlated with retweet count. Besides, the responses for desk and car detectors are not significantly correlated with retweet count. Note that the responses that are used in object-based features are not binary values indicating the presence or absence of an object. However, they measure the possibility of an object being detected in an image at any detection scale and any spatial pyramid level. Even though, ObjectBank is one of the state-of-the-art methods, the performance of the detectors might not be as accurate as desired. Besides, using only the responses of object detectors would not be sufficient for a strong detection model. Word-phrases, and attributes might be used to increase the performance of detection and thus might also improve the performance of our predictive model. When we consider the color-based features, we only focus on simple color intensity distributions over the images.

We also find that *followersFriendsRatio*, is highly correlated with the retweet count,and improves the baseline performance significantly when used with random forest model. This feature as a measure of user behavior in twitter can be used in many other predictive modelling tasks. Furthermore, for the tweets that contain a link to a web-site, the information in such sites could also be exploited to improve the accuracy of predicting retweet count.

5. CONCLUSION

In this study we focus on the task of predicting the retweet count of tweets. We focus on content- and structure-based features as well as image-based features for our predictive models. The baseline features, with content- and structure-based features, provides an RMSE of 1.743, 1.722, and 1.553 in *log* scale and 5.713, 5.599, and 4.725 in linear scale by using linear, SVM, and random forest regression respectively. Our approach in which we augment the set of baseline features with visual cues improves the prediction accuracy of the models and provides RMSE of 1.703, 1.559, and 1.297 in *log* scale and 5.489, 4.753, and 3.659 in linear scale by using linear, SVM and random forest regression respectively. Among different sets of visual cues, we find that predictive model with object-based features provide more accurate results and is more robust than low-level image features to a change in the regression technique used. The addition of visual cues increases the accuracy of predicting which tweets are most likely to be retweeted. We conjecture that the more accurate information exploited from the links in tweets, the more accurate the prediction models should perform.

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor. We would like to thank Ibrahim Uysal for his support.

7. REFERENCES

[1] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, pages 65–74, 2011.

[2] D. Boyd, S. Golder, and G. Lotan. Tweet tweet retweet: Conversational aspects of retweeting on twitter. In *Proceedings of HICSS-43*, pages 1–10, 2010.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification 2nd Edt.* Wiley-Interscience, 2001.

[4] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference.* Back Bay Books, 2002.

[5] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.

[6] L. J. Li, H. Su, E. P. Zing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[7] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[8] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Int. Conf. on Social Computing*, pages 177–184, 2010.

[9] I. Uysal and W. B. Croft. User oriented tweet ranking: a filtering approach to microblogs. In *CIKM*, 2011.

[10] J. Yang and S. Counts. Comparing information diffusion structure in weblogs and microblogs. In *ICWSM*, 2010.

[11] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in Twitter. In *ICWSM*, pages 355–358, 2010.

[12] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *CIKM*, pages 1633–1636, 2010.

[13] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern. Predicting Information Spreading in Twitter. *Comp. and Info. Sci.*, pages 1–4, 2010.