Twitter Retweet Prediction

Sean Gransee
sean.gransee@gmail.com

Ryan McAfee
ryanmcafee2014@u.northwestern.edu

Alex Wilson
alexwilson2012@u.northwestern.edu

1. Introduction

We propose a system that will learn, for an individual Twitter user, what content they should tweet that is most likely to get them the highest number of retweets.

2. Motivation

Twitter is incredibly important for personal and company branding. Currently, there are many tools for Twitter analytics, but these all simply show some pretty graphs detailing whether or not you are getting retweets. None of these tools attempt to create the link between this data and actual user content, so as to help the user tweet more effectively.

3. Data

We collected 433,354 tweets from the top 175 users on Twitter, which are used as both the training and testing data using 30-fold cross validation. This data contains approximately 2,000 tweets from each user, the retweet count of each tweet, and the time and date of each tweet.

Originally we intended to bin different kinds of users into categories such as political figure, celebrity, news/media organization, etc. This would allow us to learn more global factors that make a tweet 'good' based on the type of user tweeting. Through performing base tests on our data, we noticed very different patterns for users who would normally be grouped into the same category. Since we have plenty of data for each twitter account (over 2,000 tweets), we decided to focus on what attributes make an individual user's tweets likely to be retweeted.

3.1 Scoring the Data

To score the data, we start by creating a Theil-Sen estimator for each user's tweet history. The Theil-Sen estimator is a regression line that is robust to outliers. Figure 1 shows an example of Justin Bieber's Theil-Sen line.

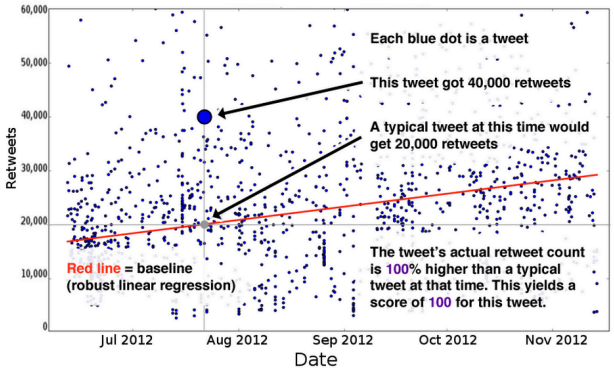


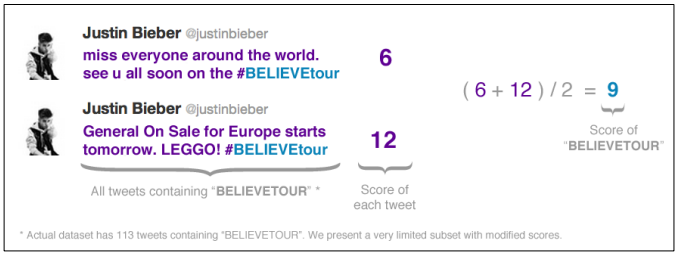
Figure 1: Justin Bieber's tweets with Theil-Sen estimator

This regression line models the increase in retweets that Justin Bieber's tweets receive as his follower base grows over time. We scored each tweet by its retweet count percentage above the Theil-Sen baseline. This scoring technique allows us to compare users with different retweet popularity as well as tweets from different times in a single user's history. As Figure 1 explains, a tweet that is retweeted 40,000 times when the Theil-Sen at this time in a user's history predicts a retweet count of 20,000 will be given a score of 100, 100% above the norm.

4. Our Learning Algorithm

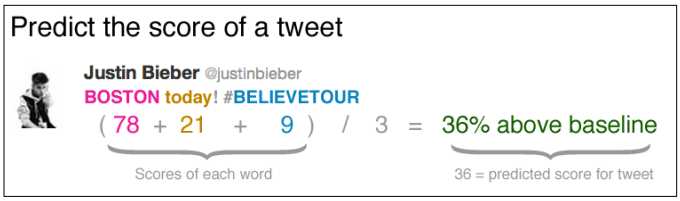
4.1 Scoring Words in a Tweet

A huge factor for determining the quality of a tweet is based on the content contained within the tweet. To extract the content attribute, we construct a dictionary from the training data to determine how powerful each word is in making a tweet retweetable. Our algorithm looks at every word in each tweet and assigns the tweet's score to the word. By averaging all of the scores a particular word is assigned, we are able to calculate a score for each word in our training set of tweets. Below is an example of calculating the score of the word "BELIEVETOUR" for Justin Bieber:



Next we apply this dictionary of word scores to the words in a new tweet. Our predictor generates a tweet score by averaging the scores of the words in the tweet.

Below is an example of how this works:



This approach inherently has an inductive bias of assuming that the words in a tweet are what make it more likely to be retweeted. We know this must have some basis in reality, because logically a user's twitter audience has words that they feel more comfortable retweeting. However, this method ignores the context behind tweets, for instance in calculating the score of Obama's "Four more years" tweet, this method doesn't work because the reason this tweet was so highly retweeted was the context of Obama getting re-elected.

4.2 Support Vector Regression

We also constructed a Support Vector Regression on the following attributes to determine what other attributes affected tweet score aside from just the words in the tweets:

- Tweet time of day
- Length of tweet (1 to 140 characters)
- Average word length in tweet
- Number of #hashtags
- Number of @mentions
- Number of links
- Time between successive tweets

We tried to combine the tweets' scores with the attribute vectors described above in a multivariate polynomial regression to learn its weight components; however, we found that simply using our text-driven analysis did a better job predicting the tweet score than any combination of these additional attributes.

To illustrate why this turned out to be the case, here is the result of analyzing just one of those attributes, time of day on our Justin Bieber example:

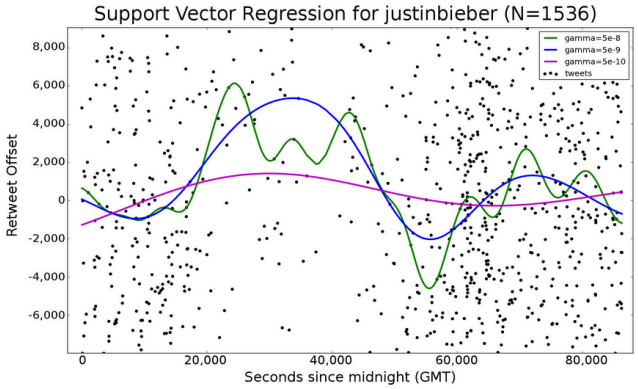


Figure 2: Support vector regression on time of day attribute

Figure 2 shows a support vector regression for three different gamma values; however, there is quite a lot of variance in all of the regression curves. Some attributes, such as "length of tweet" (not depicted in above graph), have the regression create a straight line very close to zero implying that there is nothing to be learned from that attribute. The attributes with the highest variance were given the most weight in the final predictor.

Overall, the results on these additional non-text attributes indicate that they are fairly unreliable in determining the success of a tweet. Due to Twitter's viral nature and global presence, factors like the time of day, frequency of tweets, and other non-content-driven factors have little weight in determining the retweet count. For this reason, we stuck to our content-based model for predicting retweet count.

5. Results

5.1 Evaluating Performance

We ran a 30-fold cross-validation on our data to get the mean-square error and compared it to our baseline, the Theil-Sen estimator (as described in section 3.1). We had an average improvement over the baseline of about 14%. A histogram of our percent improvement over the baseline is shown in figure 3. Since our system can more accurately predict the retweet count than the baseline, we can claim that the system learned something.

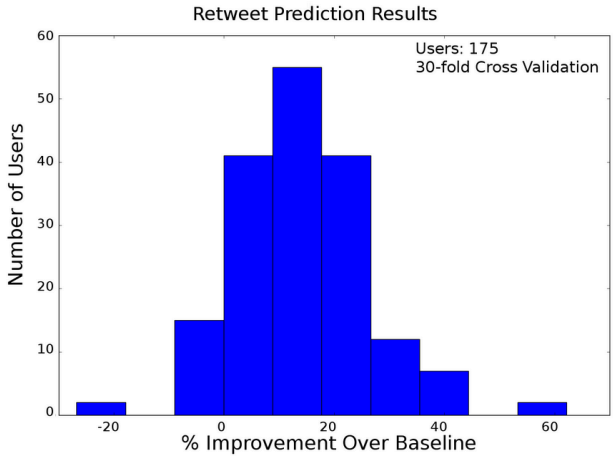


Figure 3: Histogram of improvement over baseline

6. Conclusion and Future Work

Our learning algorithm is able to predict a user's tweet 14% more accurately than the baseline prediction, which shows that our system did learn something. For some users, our system did a significantly better job predicting the tweet popularity than the baseline. This is likely because some users tweet in a more predictable fashion, and re-use many of their words across tweets, so it is easier to analyze how good those words are for that particular user.

By looking at the 'best' and 'worst' words for our users ('best' being the words with the highest score, or most likely to get the most retweets), some of these were nonsense words that were likely hashtags and their high scores come from being tweeted a single time and having a very high retweet count for the tweet they came from. This just helps to affirm our inductive bias in using our content-based algorithm. Unfortunately, trying to add other attributes that might give some context to the tweets, as shown in Section 4.2, did not help us predict the retweet score more accurately, even though it theoretically cuts down our inductive bias.

Future improvements may include analyzing pairs of words as opposed to individual words, doing sentiment analysis, and testing other semantic-driven computations on the tweet's text.

7. References

[1] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 1301-1309. http://dl.acm.org/citation.cfm?id=2145568&bnc=1

[2] Cheng, Z., Caverlee, J., and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (CIKM '10). ACM, New York, NY, USA, 759-768. DOI=10.1145/1871437.1871535 http://doi.acm.org/10.1145/1871437.1871535

[3] Sahami, M., Dumais, S., Heckerman. D., & HOrvitz, E. (1998). A Baysian approach to filtering junk e-mail. AAAI-98 Workshop on Learning for Text Categorization. Tech. Rep. WS-98-05. AAAI Press. http://robotics.stanford.edu/users/sahami/papers.html

[4] Uysal, I. and Croft, W. B. 2011. User oriented tweet ranking: a filtering approach to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (CIKM '11), Bettina Berendt, Arjen de Vries, Wenfei Fan, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.). ACM, New York, NY, USA, 2261-2264. DOI=10.1145/2063576.2063941 http://doi.acm.org/10.1145/2063576.206