Predicting Responses to Microblog Posts

Yoav Artzi *
Computer Science & Engineering
University of Washington
Seattle, WA, USA
yoav@cs.washington.edu

Patrick Pantel, Michael Gamon
Microsoft Research
One Microsoft Way
Redmond, WA, USA
{ppantel,mgamon}@microsoft.com

Abstract

Microblogging networks serve as vehicles for reaching and influencing users. Predicting whether a message will elicit a user response opens the possibility of maximizing the virality, reach and effectiveness of messages and ad campaigns on these networks. We propose a discriminative model for predicting the likelihood of a response or a retweet on the Twitter network. The approach uses features derived from various sources, such as the language used in the tweet, the user's social network and history. The feature design process leverages aggregate statistics over the entire social network to balance sparsity and informativeness. We use real-world tweets to train models and empirically show that they are capable of generating accurate predictions for a large number of tweets.

1 Introduction

Microblogging networks are increasingly evolving into broadcasting networks with strong social aspects. The most popular network today, Twitter, reported routing 200 million tweets (status posts) per day in mid-2011. As the network is increasingly used as a channel for reaching out and marketing to its users, content generators aim to maximize the impact of their messages, an inherently challenging task. However, unlike for conventionally produced news, Twitter's public network allows one to observe how messages are reaching and influencing users. One such direct measure of impact are message responses.

In this work, we describe methods to predict if a given tweet will elicit a response. Twitter provides two methods to respond to messages: replies and retweets (re-posting of a message to one's followers). Responses thus serve both as a measure of distribution and as a way to increase it. Being able to predict responses is valuable for any content generator, including advertisers and celebrities, who use Twitter to increase their exposure and maintain their brand. Furthermore, this prediction ability can be used for ranking, allowing the creation of better optimized news feeds.

To predict if a tweet will receive a response prior to its posting we use features of the individual tweet together with features aggregated over the entire social network. These features, in combination with historical activity, are used to train a prediction model.

2 Related Work

The public nature of Twitter and the unique characteristics of its content have made it an attractive research topic over recent years. Related work can be divided into several types:

Twitter Demographics One of the most fertile avenues of research is modeling users and their interactions on Twitter. An extensive line of work characterizes users (Pear Analytics, 2009) and quantifies user influence (Cha et al., 2010; Romero et al., 2011; Wu et al., 2011; Bakshy et al., 2011). Popescu and Jain (2011) explored how businesses use Twitter to connect with their customer base. Popescu and Pennacchiotti (2011) and Qu et al. (2011) investigated

* This work was conducted at Microsoft Research.

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 602–606,
Montréal, Canada, June 3-8, 2012.

how users react to events on social media. There also has been extensive work on modeling conversational interactions on Twitter (Honeycutt and Herring, 2009; Boyd et al., 2010; Ritter et al., 2010; Danescu-Niculescu-Mizil et al., 2011). Our work builds on these findings to predict response behavior on a large scale.

Mining Twitter Social media has been used to detect events (Sakaki et al., 2010; Popescu and Pennacchiotti, 2010; Popescu et al., 2011), and even predict their outcomes (Asur and Huberman, 2010; Culotta, 2010). Similarly to this line of work, we mine the social network for event prediction. In contrast, our focus is on predicting events within the network.

Response Prediction There has been significant work addressing the task of response prediction in news articles (Tsagkias et al., 2009; Tsagkias et al., 2010) and blogs (Yano et al., 2009; Yano and Smith, 2010; Balasubramanyan et al., 2011). The task of predicting responses in social networks has been investigated previously: Hong et al. (2011) focused on predicting responses for highly popular items, Rowe et al. (2011) targeted the prediction of conversations and their length and Suh et al. (2010) predicted retweets. In contrast, our work targets tweets regardless of their popularity and attempts to predict both replies and retweets. Furthermore, we present a scalable method to use linguistic lexical features in discriminative models by leveraging global network statistics. A related task to ours is that of response generation, as explored by Ritter et al. (2011). Our work complements their approach by allowing to detect when the generation of a response is appropriate. Lastly, the task of predicting the spread of hashtags in microblogging networks (Tsur and Rappoport, 2012) is also closely related to our work and both approaches supplement each other as measures of impact.

Ranking in News Feeds Different approaches were suggested for ranking items in social media (Das Sarma et al., 2010; Lakkaraju et al., 2011). Our work provides an important signal, which can be incorporated into any ranking approach.

3 Response Prediction on Twitter

Our goal is to learn a function f that maps a tweet x to a binary value $y \in \{0, 1\}$, where y indicates if x will receive a response. In this work we make no distinction between different kinds of responses.

In addition to x, we assume access to a social network $\mathcal{S}$, which we view as a directed graph $\langle U, E \rangle$. The set of vertices U represents the set of users. For each $u', u'' \in U$, $\langle u', u'' \rangle \in E$ if and only if there exists a *following* relationship from u' to u''.

For the purpose of defining features we denote x_t as the text of the tweet x and $x_u \in U$ the user who posted x. For training we assume access to a set of n labeled examples $\{\langle x_i, y_i \rangle : i = 1 \dots n\}$, where the label indicates whether the tweet has received a response or not.

3.1 Features

For prediction we represent a given tweet x using six feature families:

Historical Features Historical behavior is often strong evidence of future trends. To account for this information, we compute the following features: ratio of tweets by x_u that received a reply, ratio of tweets by x_u that were retweeted and ratio of tweets by x_u that received both a reply and retweet.

Social Features The immediate audience of a user x_u is his followers. Therefore, incorporating social features into our model is likely to contribute to its prediction ability. For a user $x_u \in U$ we include features for the number of followers (indegree in $\mathcal{S}$), the number of users x_u follows (outdegree in $\mathcal{S}$) and the ratio between the two.

Aggregate Lexical Features To detect lexical items that trigger certain response behavior we define features for all bigrams and hashtags in our set of tweets. To avoid sparsity and maintain a manageable feature space we compress the features using the labels: for each lexical item l we define R_l to be the set of tweets that include l and received a response, and N_l to be the set of tweets that contain l and received no response. We then define the integer n to be the rounding of $\frac{|R_l|}{|N_l|}$ to the nearest integer. For each such integer we define a feature, which we increase by 1 when the lexical item l is present in x_t.

We use this process separately for bigrams and hashtags, creating separate sets of aggregate features.

Local Content Features We introduce 45 features to capture how the content of x_t influences response behavior, including features such as the number of stop words and the percentage of English words. In addition we include features specific to Twitter, such as the number of hash tags and user references.

Posting Features Past analysis of Twitter showed that posting time influences response potential (Pear Analytics, 2009). To examine temporal influences, we include features to account for the user's local time and day of the week when x was created.

Sentiment Features To measure how sentiment influences response behavior we define features that count the number of positive and negative sentiment words in x_t. To detect sentiment words we use a proprietary Microsoft lexicon of 7K positive and negative terms.

4 Evaluation

4.1 Learning Algorithm

We experimented with two different learning algorithms: Multiple Additive Regression-Trees (MART) (Wu et al., 2008) and a maximum entropy classifier (Berger et al., 1996). Both provide fast classification, a natural requirement for large-scale real-time tasks.

4.2 Dataset

In our evaluation we focus on English tweets only. Since we use local posting time in our features, we filtered users whose profile did not contain location information. To collect Tweeter messages we used the entire public feed of Twitter (often referred to as the Twitter Firehose). We randomly sampled 943K tweets from one week of data. We allowed an extra week for responses, giving a response window of two weeks. The majority of tweets in our set (90%) received no response. We used 750K tweets for training and 188K for evaluation. A separate data set served as a development set. For the computation of aggregate lexical features we used 186M tweets from the same week, resulting in 14M bigrams and 400K hash tags. To compute historical features, we sampled 2B tweets from the previous three months.

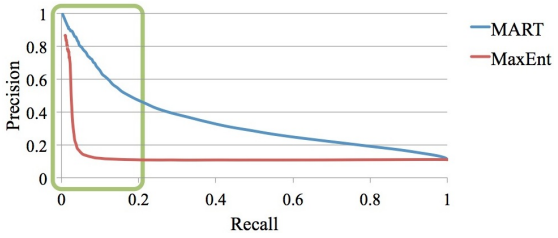

Figure 1: Precision-recall curves for predicting that a tweet will get a response. The marked area highlights the area of the curve we focus on in our evaluation.

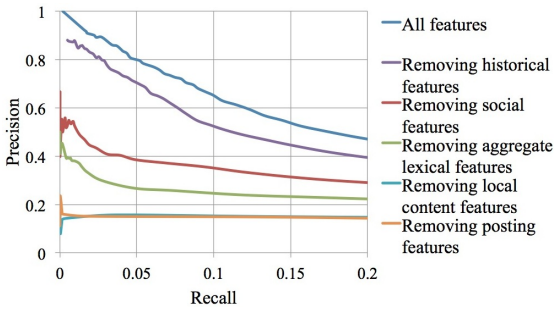

Figure 2: Precision-recall curves with increasing number of features removed for the marked area in Figure 1. For each curve we removed one additional feature set from the one above it.

4.3 Results

Our evaluation focuses on precision-recall curves for predicting that a given tweet will get a response. The curves were generated by varying the confidence measure threshold, which both classifiers provided. As can be seen in Figure 1, MART outperforms the maximum entropy model. We can also see that it is hard to predict response behavior for most tweets, but for a large subset we can provide a relatively accurate prediction (highlighted in Figure 1). The rest of our analysis focuses on this subset and on results based on MART.

To better understand the contribution of each feature set, we removed features in a greedy manner. After learning a model and testing it, we removed the feature family that was overall most highly ranked by MART (i.e., was used in high-level splits in the decision trees) and learned a new model. Figure 2 shows how removing feature sets degrades prediction performance. Removing historical features lowers the model's prediction abilities, although prediction quality remains relatively high. Removing social features creates a bigger drop in performance. Lastly, removing aggregate lexical features and lo-

cal content features further decreases performance. At this point, removing posting time features is not influential. Following the removal of posting time features, the model includes only sentiment features.

5 Discussion and Conclusion

The first trend seen by removing features is that local content matters less, or at least is more complex to capture and use for response prediction. Despite the influence of chronological trends on posting behavior on Twitter (Pear Analytics, 2009), we were unable to show influence of posting time on response prediction. Historical features were the most prominent in our experiments. Second were social features, showing that developing one's network is critical for impact. The third most prominent set of features, aggregate lexical features, shows that users are sensitive to certain expressions and terms that tend to trigger responses.

The natural path for future work is to improve performance using new features. These may include clique-specific language features, more properties of the user's social network, mentions of named entities and topics of tweets. Another direction is to distinguish between replies and retweets and to predict the number of responses and the length of conversations that a tweet may generate. There is also potential in learning models for the prediction of other measures of impact, such as hashtag adoption and inclusion in "favorites" lists.

Acknowledgments

We would like to thank Alan Ritter, Bill Dolan, Chris Brocket and Luke Zettlemoyer for their suggestions and comments. We wish to thank Chris Quirk and Qiang Wu for providing us with access to their learning software. Thanks to the reviewers for the helpful comments.

References

S. Asur and B.A. Huberman. 2010. Predicting the future with social media. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*.

E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Peoceedings of the ACM International Conference on Web Search and Data Mining*.

R. Balasubramanyan, W.W. Cohen, D. Pierce, and D.P. Redlawsk. 2011. What pushes their buttons? predicting comment polarity from the content of political blog posts. In *Proceedings of the Workshop on Language in Social Media*.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*.

D. Boyd, S. Golder, and G. Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the International Conference on System Sciences*.

M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

A. Culotta. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the Workshop on Social Media Analytics*.

C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the International Conference on World Wide Web*.

A. Das Sarma, A. Das Sarma, S. Gollapudi, and R. Panigrahy. 2010. Ranking mechanisms in twitter-like forums. In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

C. Honeycutt and S.C. Herring. 2009. Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the International Conference on System Sciences*.

L. Hong, O. Dan, and B. D. Davison. 2011. Predicting popular messages in twitter. In *Proceedings of the International Conference on World Wide Web*.

H. Lakkaraju, A. Rai, and S. Merugu. 2011. Smart news feeds for social networks using scalable joint latent factor models. In *Proceedings of the International Conference on World Wide Web*.

Pear Analytics. 2009. Twitter study.

A.M. Popescu and A. Jain. 2011. Understanding the functions of business accounts on twitter. In *Proceedings of the International Conference on World Wide Web*.

A.M. Popescu and M. Pennacchiotti. 2010. Detecting controversial events from twitter. In *Proceedings of the International Conference on Information and Knowledge Management*.

A.M. Popescu and M. Pennacchiotti. 2011. Dancing with the stars, nba games, politics: An exploration of twitter users response to events. In *Proceedings of the*

International AAAI Conference on Weblogs and Social Media.

A.M. Popescu, M. Pennacchiotti, and D. Paranjpe. 2011. Extracting events and event descriptions from twitter. In *Proceedings of the International Conference on World Wide Web*.

Y. Qu, C. Huang, P. Zhang, and J. Zhang. 2011. Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*.

A. Ritter, C. Cherry, and B. Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

A. Ritter, C. Cherry, and B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

D. Romero, W. Galuba, S. Asur, and B. Huberman. 2011. Influence and passivity in social media. *Machine Learning and Knowledge Discovery in Databases*, pages 18–33.

M. Rowe, S. Angeletou, and H. Alani. 2011. Predicting discussions on the social semantic web. In *Proceedings of the Extended Semantic Web Conference*.

T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web*.

B. Suh, L. Hong, P. Pirolli, and E. H. Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the IEEE International Conference on Social Computing*.

M. Tsagkias, W. Weerkamp, and M. De Rijke. 2009. Predicting the volume of comments on online news stories. In *Proceedings of the ACM Conference on Information and Knowledge Management*.

M. Tsagkias, W. Weerkamp, and M. De Rijke. 2010. News comments: Exploring, modeling, and online prediction. *Advances in Information Retrieval*, pages 191–203.

O. Tsur and A. Rappoport. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

Q. Wu, C.J.C. Burges, K.M. Svore, and J. Gao. 2008. Ranking, boosting, and model adaptation. *Tecnical Report, MSR-TR-2008-109*.

S. Wu, J.M. Hofman, W.A. Mason, and D.J. Watts. 2011. Who says what to whom on twitter. In *Proceedings of the International Conference on World Wide Web*.

T. Yano and N.A. Smith. 2010. Whats worthy of comment? content and comment volume in political blogs. *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

T. Yano, W.W. Cohen, and N.A. Smith. 2009. Predicting response to political blog posts with topic models. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.