Predicting Good Twitter Conversations

Zach Wood-Doughty[1,2], Prabhanjan Kambadur[1], Gideon Mann[1]
[1] Bloomberg, L.P., 731 Lexington Ave, New York, NY, 10022
[2] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD 21211
{zwooddoughty,pkambadur,gmann}@bloomberg.net

Abstract

Twitter is a primary source of social media data, but users' behaviors with *retweeting* have received much more study than have users' interactions with *replying*. To explore the substantial proportion of all tweets which are posted in reply to another tweet, we collected a dataset of millions of Twitter conversations and examined the conversational dynamics between users. We built models to predict whether a tweet will start, end, or maintain conversations, and whether a tweet will receive high- or low-quality replies.

1 Introduction

Twitter data is widely used in social science research to gain real-time insights into the opinions and behaviors of millions of people around the world. While many users rely on Twitter primarily for news and "status updates," the platform also hosts millions of discussions between its users (Kwak et al., 2010). Most Twitter posts are public, allowing any user to reply to them and begin a discussion. Posts from popular accounts often receive thousands of replies, and many users engage with one another in constructive conversations despite the content limitations of the Twitter platform.

To study conversational dynamics, we used Gnip and the standard Twitter API to collect a dataset of 15.5M tweets containing 1.5M conversations from a three-day period in May 2018. This is one of the largest dataset of Twitter conversations constructed for research purposes Ritter et al. (2010).

We labeled our dataset using distant supervision to assess the *engagement* and *quality* of conversations. We operationally defined a tweet to be engaging if it receives replies (Torres et al., 2017). We are particularly interested in turn-taking dynamics: replies from user B to user A which result in user A replying in turn to user B. To label tweets for their quality, we used a large annotated dataset of Wikipedia Talk comments which are human-annotated for aggression, toxicity, or 'personal attack' (Wulczyn et al., 2016). We trained a model to predict a continuous score for each of these three annotations, and then used that model to infer those labels for all tweets in our conversational dataset.

We then trained models to predict whether a given tweet will receive replies and whether those replies will be of high or low quality, as defined by our noisy labels. We used simple neural models trained on features learned from three modalities: the tweet text, the author's profile, and the reply-structure of the discussion so far. We studied to what extent these different feature types contribute to predicting replies. We clustered the representations learned from these models to explore broad trends in Twitter conversational behavior and to show how the conversations we studied differ from data collected in past work.

References

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *WWW*, pages 591–600.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *NAACL*, pages 172–180.

Johnny Torres, Carmen Vaca, and Cristina L Abad. 2017. What ignites a reply?: Characterizing conversations in microblogs. In *BDCAT*, pages 149–156.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Wikipedia Detox.