Dan Saunders

Estimating and Explaining Twitter Engagement

For this project, I will use a dataset of tweets gathered from the public Twitter API (https://dev.twitter.com/docs). Specifically, I will gather tweets from either influential politicians or celebrities along with accompanying metadata (likes, retweets, replies, time posted, hashtags). The aim of the project is to estimate the number of likes, retweets, or replies a tweet by a certain Twitter account with gather (a regression problem), as well as to understand the most important variables underlying tweet engagement via dimensionality reduction.

For preprocessing, after obtaining the data, I will need to remove tweets that might not be relevant (such as quoted or retweeted tweets). It will then be useful to select features from each datapoint that might be useful in the model-training phase, such as a bag-of-words or parts-of-speech representation of the tweet text, a list of the hashtags, and the time it was posted. Depending on the choice of regression value(s), likes, retweets, or replies may also be used as part of the training phase. Feature selection will be used to investigate the most predictive variables. I will begin investigation with simple models such as linear regression and linear support vector machines, and investigate how results improve with applying random forests and neural networks. For the dimensionality reduction task, I plan to filter out words that score low on information theory statistics, measure distance between tweets using cosine similarity, and use the method of latent dirichlet allocation to representation tweets as a mixture of topics. Generally, I will use the sciki-learn Python software package unless additional methods are needed, in which case I will implement them myself or find another package which contains them.

I plan to find a small number of “celebrity” Twitter accounts, from which I will collect several thousand tweets each. I will apply the pipeline individually to each account, and study how the models perform across the different subjects, and eventually train another model which will predict regression values for as-yet unseen accounts. For this, additional training parameters may be necessary (such as number of followers / following). The dimensionality reduction methods will be evaluated based on the degree to which they explain the datasets and make the data easier to understand visually or with simple language.

Reference [1] explores the effectiveness of certain tweet features for estimating the quantity of retweets, and specifically addresses the “goodness” of a tweet by a particular author. Preprocessing and feature selection on the Twitter datasets is explored in-depth. Reference [2] is an attempt to categorize what presidential candidates are talking about in their tweets. This article will be useful as more guidance for feature selection from Twitter datasets. Reference [3] explores the engagement of blog posts by using either principal components analysis or a sparse autoencoder to do dimensionality reduction, and uses either linear regression or regression trees to estimate the number of comments the blog post will gather. Reference [4] examines feature selection on and clustering of potential customers based on a similarity measure between users’ tweets. Reference [5] addresses the inference of Twitter user attributes (such as political orientation or ethnicity) by learning from user behavior, network structure, and the content of the user’s tweets.

References

[1] Estimating Effectiveness of Twitter Messages with a Personalized Machine Learning Approach (http://cs.fit.edu/~pkc/theses/sun15.pdf)

[2] Using machine learning to classify presidential candidate social media messages (http://towcenter.org/using-machine-learning-to-understand-real-time-presidential-candidate-social-media-messages/)

[3] Unsupervised Learning for Effective User Engagement
(http://web.stanford.edu/~thaipham/papers/STATS_306B_Project.pdf)

[4] Clustering a Customer Base Using Twitter Data (http://cs229.stanford.edu/proj2015/310_report.pdf)

[5] A Machine Learning Approach to Twitter User Classification
(file:///home/dan/Downloads/2886-14198-1-PB.pdf)