# Performing Analysis on Twitter Data

Animesh Patni, Chandana Ravindra Prasad, Alex Szilagyi

**Problem Overview: Describe the problem you are solving. State it as precisely as you can.**

To perform analysis of future tweet traffic for a specified set of users based on sentiment analysis of that user's historical tweets. This problem requires metrics and categorization of historical tweets to be determined based on sentiment analysis to allow for a baseline comparison of tweets before the analysis is performed. In addition to sentiment analysis, it is important to recognize other outstanding tweet features that may be affecting the retweet data.The uniqueness of this project lies in the ability to perform analysis by examining classification of a user's tweets.

**Data: Which data are you using; how did you collect it?**

We have used Tweepy (a python library to utilize the twitter API) to gather data from 12 different user timelines. Two users were selected to represent a respective field, one male and one female. The 12 users that were selected are as follows: ['BarackObama', 'HillaryClinton', 'KingJames', 'serenawilliams', 'katyperry', 'justinbieber', 'jk_rowling', 'johngreen', 'SpaceX', 'pepsi', 'elonmusk', 'mtbarra']. These users were selected based on activity to ensure that enough data could be gathered from their respective user timelines.

By selecting the users in this manner, it is possible to limit any bias that may be put on the dataset. It is also possible to gather additional metrics since we can now categorize the users. By further understanding where the data is coming from, it is possible to gain a deeper understanding of the results without limiting the scope of our data.

To grab the data, we utilize Tweepy, a python library that allows us to make twitter API calls. The API call that we reference is user_timeline, a call that allows us to gather 200 tweets from a specific user. Due to twitter rate limiting, this call had to be repeated until we obtained all of the data. Twitter stores up to 3,200 of the most recent tweets on a user's timeline. The first 200 tweets were gathered and stored, of which we utilized the last tweet's #ID to gather the next 200 tweets older than #ID.

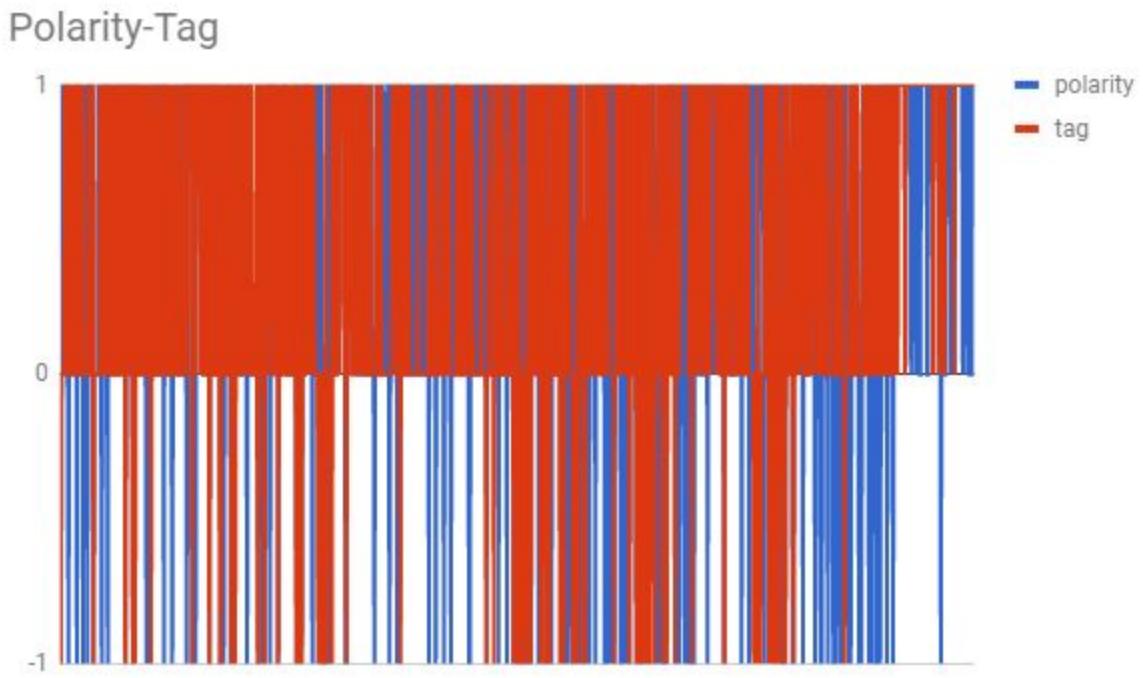
**Method: What method or algorithm are you using. Are you using an existing library to do so? Did you introduce any new variations to these methods? How will you evaluate the results? Which baselines will you compare against?**

Once this data was gathered, existing libraries were utilized to perform sentiment analysis and begin training our model. The library that we used to perform the preliminary semantic analysis was TextBlob. The polarity of each tweet was returned and marked based on their overall sentiment analysis. Upon observing this automated process, we recognized that the outcome was unsuitable due to its automated nature. In order to obtain a higher accuracy of classifications, each tweet was manually tagged. This manual tagging allowed for a baseline to be generated for a user's average retweet count per sentiment classification.  In the next section

we list some outlier analysis of this process. These mostly stem from failures in negation detection, adjective, and adverbs. In order to fix this issue, we propose that we further utilize the NLTK package to utilize features such as PoS tagging. We plan to make a comparison between Logistic Regression, Naive Bayes, and SVM to ensure that the greatest accuracy is obtained before proceeding with the analysis.

**Intermediate/Preliminary Experiments & Results: State and evaluate your results up to the milestone**

Through our preliminary analysis of semantic analysis on the 1,200 tweets we have gathered,




--

Outlier analysis:

| | | | | | |
|---|---|---|---|---|---|
| # # # | b'RT @lumos: Institutionalisation of unaccompanied migrant & refugee children denies them their right to a family & puts them at greater risk\xe2\x80\xa6' | 4 | 190 | 0 | -4 |

---- “right” ---

| | | | | | |
|---|---|---|---|---|---|
| # # # | b"RT @hugorifkind: This tweet wouldn't have happened five years ago. How have we let these proudly racist rats crawl out of our national wood\xe2\x80\xa6" | 4 | 1824 | 0 | -4 |

-------”proudly”-------

| | | | | | |
|---|---|---|---|---|---|
| # # # | b'RT @qikipedia: Adding iron oxide to putty produces magnetic putty, which can exhibit some quite interesting behaviour. See the full video h\xe2\x80\xa6' | 4 | 1187 | 0 | 0 |

---------”interesting behaviour”- not sentimental - ambiguous------

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| # | # | b'RT @qikipedia: People with red hair are less responsive to anaesthetic.' | -4 | 707 | 0 | 0 | |

---------just a fact “less responsive”---not neg-----

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| # | # | b'RT @Nick_Pettigrew: Yes, but all 67 million of us won\xe2\x80\x99t fit in Parliament. That\xe2\x80\x99s why we have MPs to represent us. Stop me if this is getti\xe2\x80\xa6' | 4 | 39 65 | 0 | 0 | |

------”yes” is misleading-----

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| # | # | b'@RakieAyola @BroadwayWorldUK @HPPlayLDN @NimaxTheatres Congratulations Rakie! So well deserved! \xf0\x9f\x8d\xbe\xe2\x9d\xa4\xef\xb8\x8f\xf0\x9f\x91\x8f' | 0 | 28 | 73 | 4 | |

------”congratulations” should’ve been positive-----

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| jk_rowling | # | b'Thread \xf0\x9f\x8c\x9e\xe2\x9d\xa4\xef\xb8\x8f\xf0\x9f\x8c\x99 https://t.co/iZv7sjjHzj' | 0 | 10 09 | 81 77 | 0 | |

-------dealing with only links------

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| johngreen | # | b'@smartereveryday I\xef\xb8\x8f want to watch a game at your house!! Looks fun!!' | -4 | 2 | 13 5 | 4 | |

-------”looks fun” marked as neg-----

Challenges:
What assumptions can be made during the training phase? How do we ensure that the training data doesn't have biased information which leads to incorrect/ biased analysis?

How to identify trade-offs that are to be made wrt accuracy?

Given the size of our dataset to be analysed are there any optimization techniques we should be looking at?

**Related work: Summarize at least five related research papers related to your project. How is your project similar/different?**

Prediction of retweet cascade size over time; Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, Andrey Kustarev, Yandex Leo Tolstoy st. 16, Moscow, Russia
https://dl.acm.org/citation.cfm?id=2398634

The authors of this paper provide an analysis of retweet cascades. The idea behind this topic is to predict the size of the cascade of the tweet at some future time. This is similar to ours in the notion that we wish to examine the spread (or retweet) of a single user's tweet. In addition, this paper mentions some features of a tweet such as a user's social network (number of friends, followers, ect..), content features (the length of the tweet, it's semantic analysis, ect..) as well as a few others. Our paper differentiates from this in the fact that we are not interested in examining a large number of features. This paper is still impactful since it is important to recognize that these features are play a large role in analyzing the true impact of a retweet.

Analyzing User Retweet Behavior on Twitter; Zhiheng Xu, Qing Yang
Institute of Automation, Chinese Academy of Sciences
https://dl.acm.org/citation.cfm?id=2457094

In this paper, the authors provide an analysis of user retweet behavior on twitter. They do this by providing an analysis of retweet behavior and train prediction models on different classification frameworks. This is very similar to our current step, in which we wish to analyze different classification frameworks as mentioned above to ensure a high accuracy is obtained. This paper primarily focuses on reflecting a variety of features from the user and reflects the importance of a variety of features. Instead of comparing these features, we wish to perform analysis utilizing semantic values for tweets.

Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network; Bongwon Suh, Lichan Hong, Peter Pirolli, Ed H. Chi
Palo Alto Res. Center, Inc., Palo Alto, CA, USA
http://ieeexplore.ieee.org/document/5590452/

This paper provides an analysis of tweet features and their correlation to retweets. Some of these features include the number of URLs in a tweet, the number of followers, hashtags, days, status, and favorites. They have found that URLs and hashtags correlate highly with retweetability. Although this paper is fundamentally different than ours, it provides a frame of reference that must be recognized. In order to eliminate any bias in our data and ensure that the attention that we are trying to measure correlates a tweet's semantic analysis, we must be open to recognizing any additional features that play a large role in the correlation of retweets.

Understanding Email Writers: Personality Prediction from Email Messages Jianqiang Shen, Oliver Brdiczka, and Juan Liu Palo
Alto Research Center, 3333 Coyote Hill Road,
Palo Alto, CA 94304, USA {jianqiang.shen,oliver.brdiczka,juan.liu}@parc.com
http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00049?class=ref+nowrap+pdf

In this paper, they infer the personality of users based on the content of their emails. Such inference can be used for various applications such as better personalization, recommendation, and targeted advertising. They have a way to deal with the private and sensitive nature of email content without barging in on the privacy. Based on the Big Five personality model, they have trained predictors to work on extracted email features. They report prediction performance of 3 generative models with different assumptions. And the results show that personality prediction is feasible, and the email feature set can predict personality with reasonable accuracies.

The authors speak about two approaches for developing their SOCAL(Sentiment Orientation CALculator), one being lexicon-based approach and text classification approach being the other one. The dictionary for lexicon-based approach is created manually with the focus on adjectives being used as indicators of the semantic orientation.

Twitter Sentiment Analysis; Sarlan, A.; Nadam, C.; Basri, S.
The authors in this paper examine sentiment analysis of web based applications and the limitations that are faced. These limitations appear in the use of inappropriate English (slang terms) as well as the limitations placed upon a text-only message, which are void of all facial expressions and body language. This paper provides groundwork for a lexicon-based and machine learning based approach of semantic analysis. This groundwork supports our paper but does not carry out any technical implementations.

**Who does what: State which group member is responsible for which aspects of the project.**
Thus far, Animesh has worked with the collection of data through the TwitterAPI calls and the TextBlob polarity. Chandana was responsible for documenting the outlier polarities and Alex was responsible for researching related works and documentation of the progress. Alex and Chandana have been taken part in the manual tagging of the polarity of tweets. Each member will participate in the phase of testing accuracy for each model (3/25 in timeline). Each member is given a framework to analyze as follows. Animesh is responsible for Naive Bayes. Alex is responsible for LR. Chandana is responsible for SVM.

Once this phase is complete, the team will come together for the 4/1 deadline and perform analysis on these results to work on the model. Lastly, the team will ensure that each member participates in the formation and writing of the paper and presentation.

**Timeline: What are the remaining steps you plan to complete, and when do you plan to complete them?**

3/25 - Ensure raw data is collected and categorized for test set. Test accuracy of LR, naive bayes, svm to allow for a comparison to be made between the models.
4/1 - Utilize the highest accuracy model to perform analysis on the test set of tweets.
4/9 - Future Works: Look into further categorizing users/tweets
Presentation is due on 4/15. Report is due on 4/18.

**References: list of references cited in your report.**

Prediction of retweet cascade size over time; Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, Andrey Kustarev, Yandex Leo Tolstoy st. 16, Moscow, Russia (https://dl.acm.org/citation.cfm?id=2398634)

Analyzing User Retweet Behavior on Twitter; Zhiheng Xu, Qing Yang Institute of Automation, Chinese Academy of Sciences (https://dl.acm.org/citation.cfm?id=2457094)

Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network; Bongwon Suh, Lichan Hong, Peter Pirolli, Ed H. Chi Palo Alto Res. Center, Inc., Palo Alto, CA, USA (http://ieeexplore.ieee.org/document/5590452/)

Understanding Email Writers: Personality Prediction from Email Messages Jianqiang Shen, Oliver Brdiczka, and Juan Liu Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA {jianqiang.shen,oliver.brdiczka,juan.liu}@parc.com (http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00049?class=ref+nowrap+pdf)

Twitter Sentiment Analysis; Sarlan, A.; Nadam, C.; Basri, S. (http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=7066632)