**Social Propagation of Twitter**

**Nina Dinh, Muftiat Ogunsanya, Claire Pitman**

**May 10th, 2016**

**ABSTRACT:** This project studies the spread of tweets in four social movement related hashtags: #BlackLivesMatter, #FeelTheBern, #MakeAmericaGreatAgain, and #LoveWins. For each of these hashtags, 100 tweets were analyzed, considering the factors of sentiment, number of followers, and number of followers who also tweeted with the same hashtag in a twelve hour time frame. This information was used to determine if a prediction could be made on how many users will tweet with the same hashtag using factors of sentiment and number of followers through machine learning. The machine learning results were inconclusive. For the linear regression model, the machine predicted values ranging from -11.09 to 38.46, and for the SVM regression model, the machine was only able to predict values from 0.0754 and 2.96.

**Keywords:** sentiment; social movement; Twitter; pattern prediction;

**1 INTRODUCTION**

While Twitter began as a social media platform, it has now become a contemporary means for social activism towards causes such as the Arab Spring and Occupy movements. Serving as a hub for people to express their opinions about such controversial issues, people are able to connect and spread their voices by simply tacking on “#ArabSpring” or “#OccupyWallStreet” onto their end of their tweets. As social activism moves from physical space to cyberspace, with the “emergence of new forms of protest,” we wanted to look more closely at what exactly influenced the flow and spread of social movements online [2]. We studied the spread of tweets in various social movements, focusing specifically on how factors such as a user’s tweet sentiment and number of followers can predict the number of that user’s followers who will tweet with a specific hashtag. For example, if a Twitter user posts a negative tweet towards an issue, how likely is it that his/her followers will do the same, and what factors will cause it to become more likely? While “URLs that were rated more interesting and/or elicited more positive feelings by workers on Mechanical Turk were more likely to spread,” we are looking at more specifically controversial environments within Twitter rather than general website environments [1]. For our methodology, we first chose four popular hashtags dealing with social movements and obtained thousands of tweets and tweet details for each hashtag. After preprocessing the data and using sentiment analysis, we evaluated the data through machine learning. For the linear regression model, the machine predicted values ranging from -11.09 to 38.46, and for the SVM regression model, the machine was only able to predict values from 0.0754 and 2.96.

**2 METHOD**

The approach to solve the problem proposed in this project began with choosing popular trending topics. In this project, we focused on four hashtags: #BlackLivesMatter, #FeelTheBern, #MakeAmericaGreatAgain, and #LoveWins. #BlackLivesMatter was the top trending topics of 2015 and has still been a very active hashtag so far in 2016. In addition, Donald Trump and Bernie Sanders are two political candidates with strong presences on social media. The slogan for each of their campaigns have been widely used from the end of 2015 into 2016 on many social media platforms, specifically 2016. #LoveWins was a very popular hashtag after the Supreme Court decision that legalized same sex marriage and has recent increased in popularity due to a recent Old Navy advertisement. Because these hashtags continue to be popular and widely used on Twitter, it seemed best to focus on these hashtags for this project.

The next phase of the project was getting usable datasets. Instead of searching a dataset online, we decided to curate tweets using the Twitter Archiver extension for Google Sheets. The Twitter Archiver extension gathers the date and time of the tweet, the screen name of the user, the published name of the user, the actual text of the tweet, the tweet ID, app used, the user’s number of followers and follows, the number of retweets and likes the tweet received, and other user information such as location and bio. The Twitter Archiver was used to curate tweets that only included these specific hashtags. We began ran the Twitter Archiver from March 19, 2016 to March 28,2016 for #FeeltheBern and #MakeAmericaGreatAgain, from March 17, 2016 to April 11, 2016 for #BlackLivesMatter, and on May 6, 2016 for #LoveWins. After collecting the tweets for each hashtag, we had 17,453 tweets for #FeelTheBern, 7673 tweets for #MakeAmericaGreatAgain, 43,036 tweets for #BlackLivesMatter, and 3,080 tweets for #LoveWins.

After collecting the tweets, preprocessing had to occur. From each tweet, we wanted the screen name, tweet, number of retweets, number of followers, number of followers who tweeted with the same hashtag, and sentiment. In order to conduct the preprocessing, we used the Tweepy API, Sentiwordnet, and the natural language toolkit (NLTK). The sentiment and number of followers who tweeted with the same hashtag was only collected for each of the first 100 tweets because of time constraints. However, for the number of followers who tweeted with the same hashtag took into account the tweets/users from the first 12 hours of each dataset. Tweepy was used to get the user ID and follower IDs for each user in a dataset in order to calculate how many of an user’s followers tweeted with the same hashtag. Sentiwordnet and NLTK was used to calculated the sentiment of the first 100 tweets. The final dataset for each hashtag included the screen name, tweet, number of followers, number of retweets, sentiment, and the number of followers who tweeted with the same hashtag.

The final dataset for each dataset was used for evaluation through machine learning. We were only able to use three features: time, number of followers and sentiment of the tweet text because the username/id only added more noise and the number of retweets was to close to what we're trying to predict. We also normalized the time so that the time used for training and prediction was the time since the first collected tweet. We attempted using the absolute value of sentiment however the results were worse than with raw sentiment and are omitted. We chose to use #FeelTheBern as the hash tag that would be predicted. The other three hashtags were used as training data.

For the machine learning portion we used the sklearn library because it is a well documented and widely used library. We decided that this was a regression problem because the predictions should be continuous. We then experimented with different methods that are used for regression to try to reduce the residual sum of squares. In the results section we will cover only two of these methods: linear regression, and svm based regression.

**3 EXPERIMENTAL RESULTS**

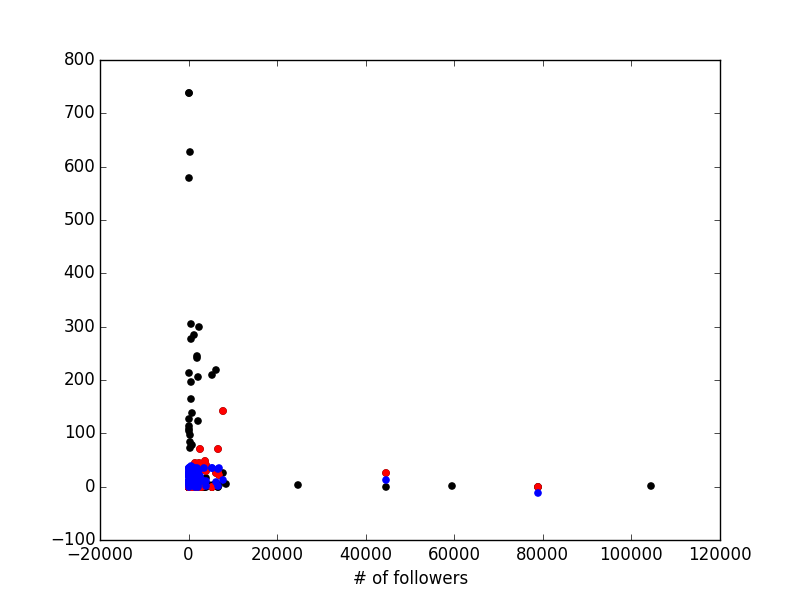
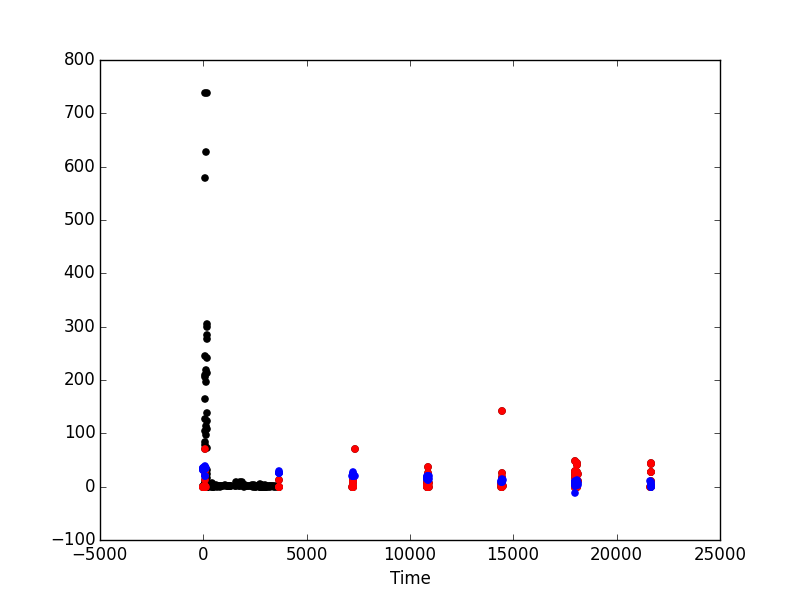
For each model we included the residual sum of squares that we were able to achieve and three graphs. Each graph shows the training data in black, the predictions for the test data in blue, and the actual labels for the test data in red. For each graph the y-axis is the number of followers that used that hashtag.

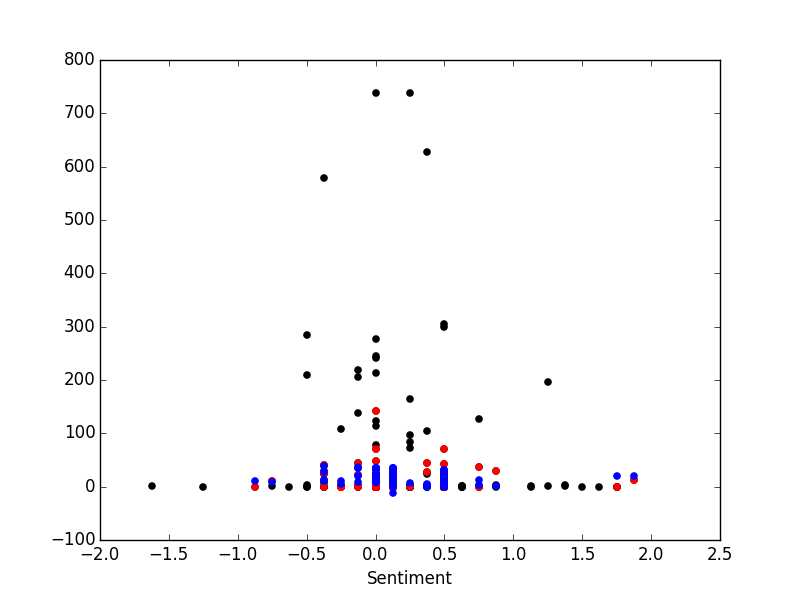
**Linear Model**

This model has a higher mean residual sum of scores but predicted a larger variance of numbers, ranging from -11.09 to 38.46.

Mean Residual sum of squares: 628.20

Variance score: -0.54



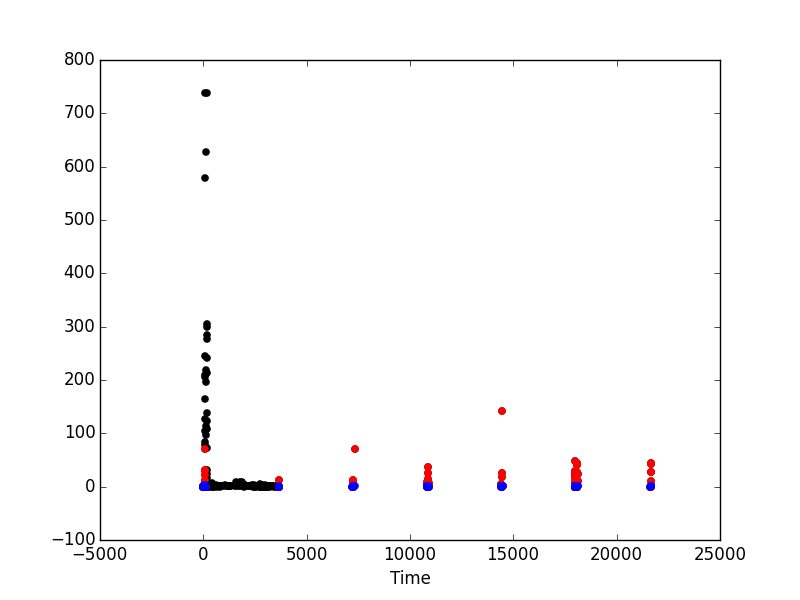
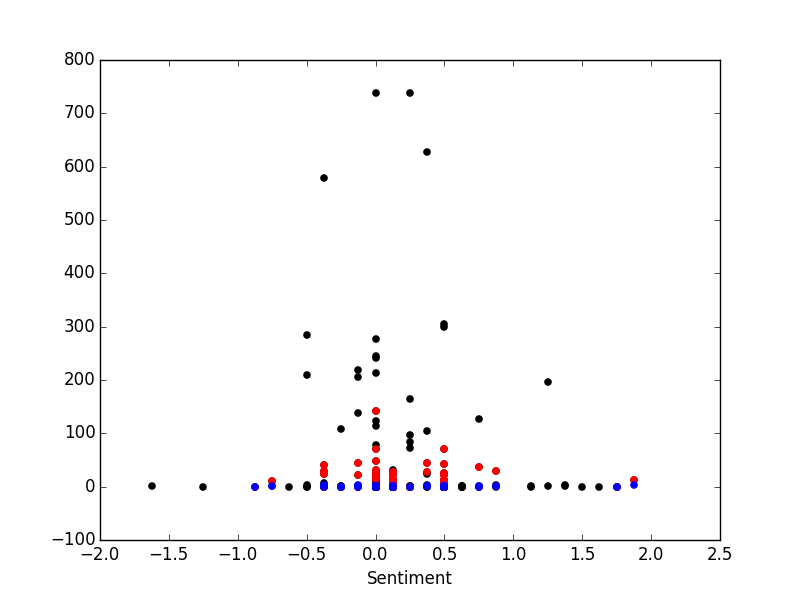
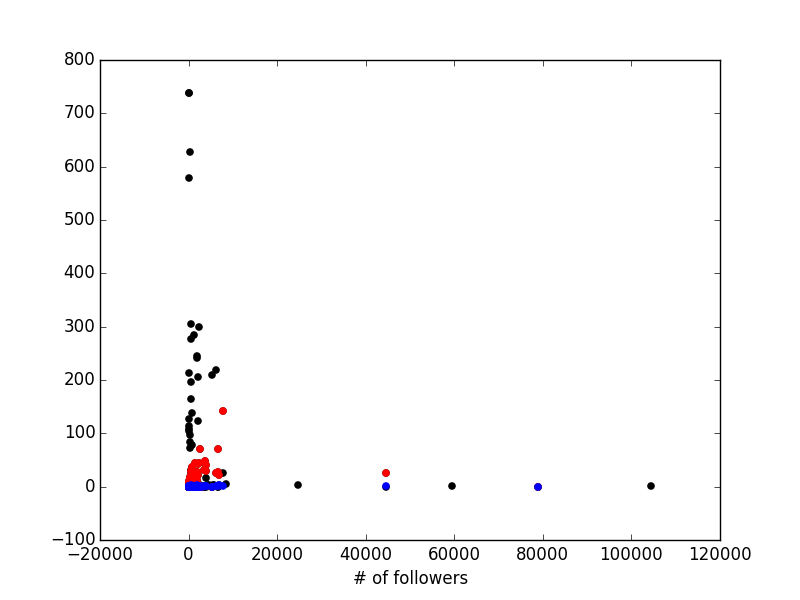


**SVM based Regression**

This model was able to achieve a lower residual sum of squares but only predicted values between 0.0754 and 2.96.

Mean Residual sum of squares: 493.45

Variance score: -0.21



The results from the machine learning were inconclusive. The linear regression attempted to make predictions, but the predictions were incorrect for the most part. While on the other hand, the SVM regression was only able to make predictions in a very small range. In addition, we observed a power law distribution for the number of followers who retweeted with the same hashtag as a function of both time and number of followers. However, there is a regular distribution for the number of followers who retweeted with the same hashtag as a function of sentiment.

**REFERENCES**

[1] Bakshy, Eytan, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. "Everyone's an Influencer." *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining - WSDM '11* (2011): 65-74.

[2] Gerbaudo, Paolo. *Tweets and the Streets: Social Media and Contemporary Activism*. N.p.: Pluto, 2012. Print.