**Influence of social media: Does your participation on social media effects your business?**

Elli Rodin(alyona.rodin@gmail.com)

**Abstract**

The role of social media is growing in our everyday life. Social media becomes integral part ofour living to self-express and to be heard. However, our participationon social media can affect our image positively aswell as negatively, depending from the publicopinion. This project will explore the possibilities of daily tweetinginfluence on the stock returns of the company. As the stock price is effected by the growth of the company, and company’s expectations. What influences the company’s growth and how we recognize the expectations? The paper will look at the tweets posted by Elon Musk (EM), using Vader NLTK to see if his messages effected the growth of his companies such as Tesla. Specifically, we examine his tweets and see if there is a correlation between his opinions on subject matter and Tesla’s return through these years.

**Keywords:** Machine Learning; Tweeter; tweet; return; Tesla; SpaceX, Sentiment Analysis, NLP

I. INTRODUCTION

At the beginning of the project, we look at the tweets which were posted by EM. We define how actively he participates on the social media – how often he posted tweets per day, how often he retweets the messages, what words the tweets contain the most, and also perform sentiment analyses to determine if those tweets were positive, negative, or just neutral. Given the information about the tweets as well the returns on the Tesla stocks we try to identify if the tweet posting indeed effect the prices and accordingly the return, the growth of the company.

II. DATA COLLECTION AND PREPROCESSING

*A. Data Collection*

The data comes from the Kaggle [1] in the .csv format. It contains the 3218 tweets over the period of 2012 to 2017 years. Another dataset which was used – information about the stock Tesla (ticker: TSLA) was downloaded from Yahoo/Finance [2].

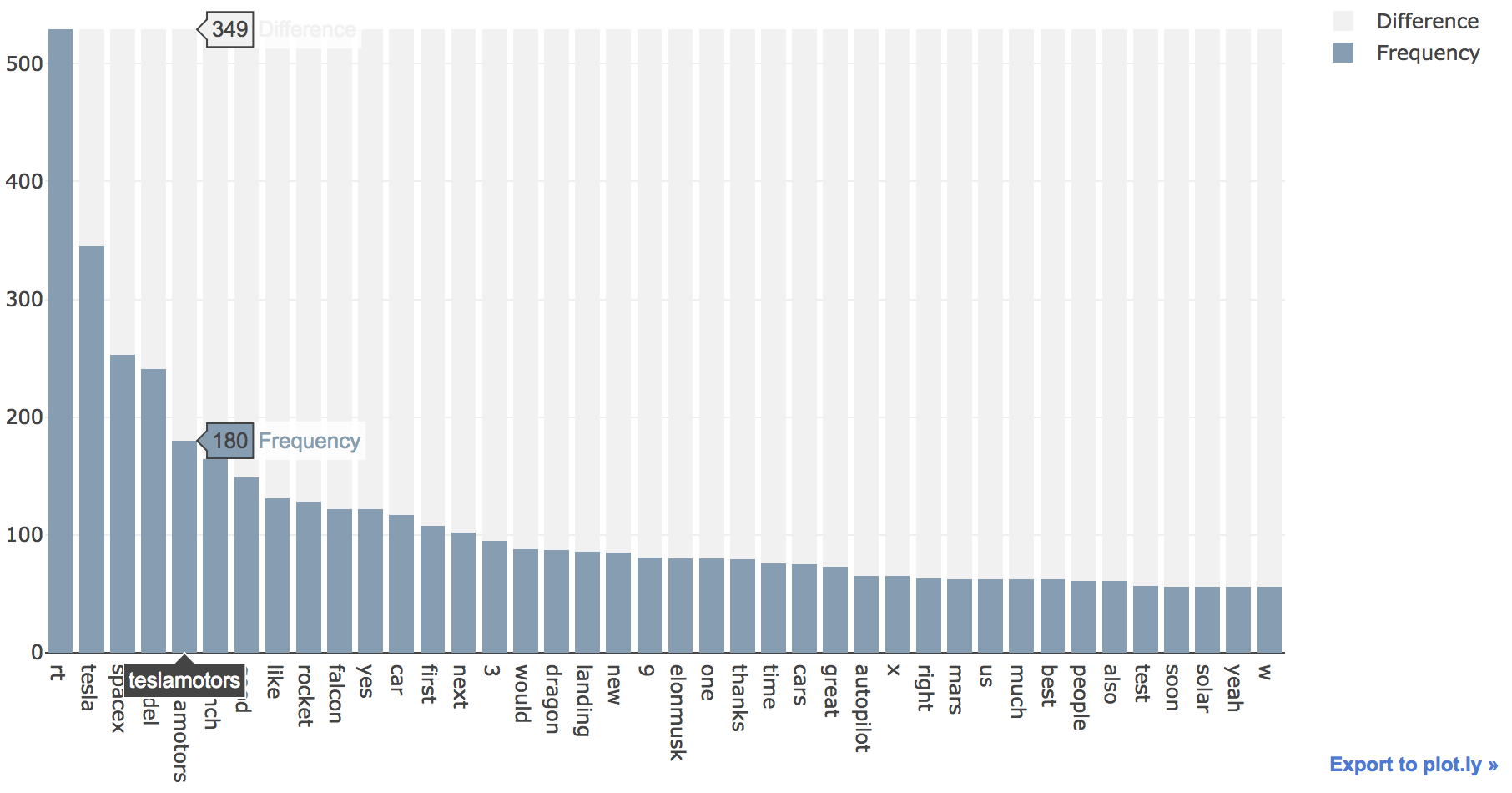
*B. Data Preprocessing*

The data on the stock prices is not complete as stock markets does not work during weekends and tweets can be posted during that period of time as well holidays. When we concatenated two datasets we removed the rows with missing data. However this technique has its disadvantages, for example it can make a reliable analysis impossible. Another way to deal with missing data is imputation (for future consideration).

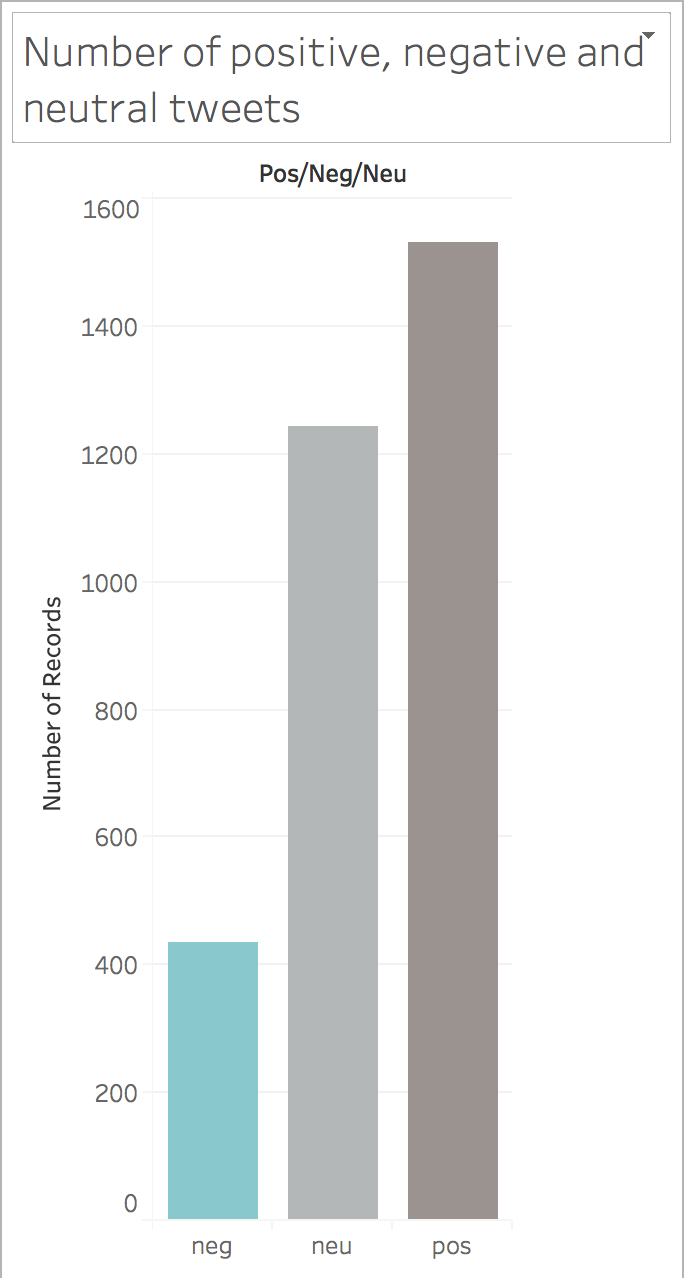
Tweets contain a lot of unnecessary information. For preprocessing we used: Tokenization (split the tweets into separate words), Stopword (removed the words such as prepositions, conjunctions, etc., removed the characters.

From the sentiment analyses we find out that the EM retweets often, and derived the most often used words which are Tesla and SpaceX.

Performing sentiment analyses of the tweets we separated each tweet onto positive, negative, and neutral. As well we added how many time he tweets per day. We notice that his participation on the tweeter through the years increased (obviously) (Pic.1), but the stock price is increasing also in the similar pattern.

****

For the stock TSLA we calculated the return and the direction of return (Up/Down). The direction we will try to predict based on the available features we had or extracted.

****

*C. Features*

There are raw features and derived features in both database. Besides that we have numerical and categorical data.

Numerical data such as Volume and AdjClose are not at the same scale as the rest data. To bring it into scale we used z-score standardization as we use logistic regression and it will be easier to learn weights. Also from the analyses of the data, we noticed that we have outliers and standardization makes the algorithm less sensitive to the outliers.

For the categorical data (nominal) we derived the dummy variables (Pos/Neg/Neu), Rt, Tesla, SpaceX, Model and Direction itself). We removed from each column one value to eliminate Multicollinearity.

III. METHODS AND ANALYSES

*A. Feature Selection*

For the feature selection we use L1 regularization. Penalizing model with strong regularization parameter we receive more features that are zeroes. We chose the regularization parameter .

Training accuracy: 0.996

Test accuracy: 0.9939

However if we decrease the regularization parameter, C=0.6, the Test accuracy increases to 0.995

Besides that I notice that if we use the absolute value of return feature the training and test accuracy drastically fall to about 0.58 and 0.56.



The most important features that it selected for us is the mood of the tweet, rt, tesla, spacex.

What we noticed here also that once we add the number tweets into the features, it shows us it significant feature too.

*B. Assessing Features*

To assess the features we use the random forest.

The features will be ranked according its importance. Indeed, the mood of the tweet is most important feature that affects the return and #tweets.



1) pos/neg/neu\_neg 0.941025

2) pos/neg/neu\_pos 0.036384

3) #tweets 0.017185

4) rt\_1 0.001385

5) tesla\_1 0.000993

6) spacex\_1 0.000803

7) model\_1 0.000769

8) list\_adjCl 0.000754

9) list\_return 0.000703

*C. Logistic Regressions*

To perform the logistic regressions, we partitioned the dataset in training and testing sets which is 70:30.

In the logistic regressions the purpose is to minimize the cost function – Sum of Squared Errors.

RESULTS

To assess the performance of the algorithm we use the confusion matrix. The confusion matrix provides the counts of the true positive, true negative, false positive, and false negative predictions of a classifier.

We can see that there is 3 total mistakes. Our model correctly classified 279 of the samples of Down (class 0) and 379 of the samples of Up (class 1)



However, the model incorrectly misclassified 2 samples from class 0 and 1 sample from class 1.

CONCLUSION

In this paper we saw that there is a dependence of return (Up/Down) from the tweets which EM posts, or simply retweets, which contains words related to his company or him. Particularly the return depends of the “mood” of the EM tweets (be it positive or negative). Based on the sentiment analyses of the tweets EM posts more positive messages and uses positive words related to his company. Thus we see that we can increase the performance of the business. But we can be easily deceived also too see the true picture.

FUTURE WORK

There are a lot of areas that can be explored yet. Below are some suggestions that will be performed in the future.

* Try other algorithms, for example Q-learning algorithm (Reinforcement Algorithm), SVM, etc.
* We used data from which deleted a lot of rows from the dataset of EM tweets which contain NaN values. Try to use mean imputation (interpolation technique).
* Include more features into the dataset, such as time the tweet was posted.
* Add sentiment analyses of tweets used by other people regarding EM or his companies and extend our question.
* Add StockTwits [3] (social media designed specifically for investors, entrepreneurs).
* Remove outliers.
* Also would be great to explore the phycological influences of social media.

REFERENCES

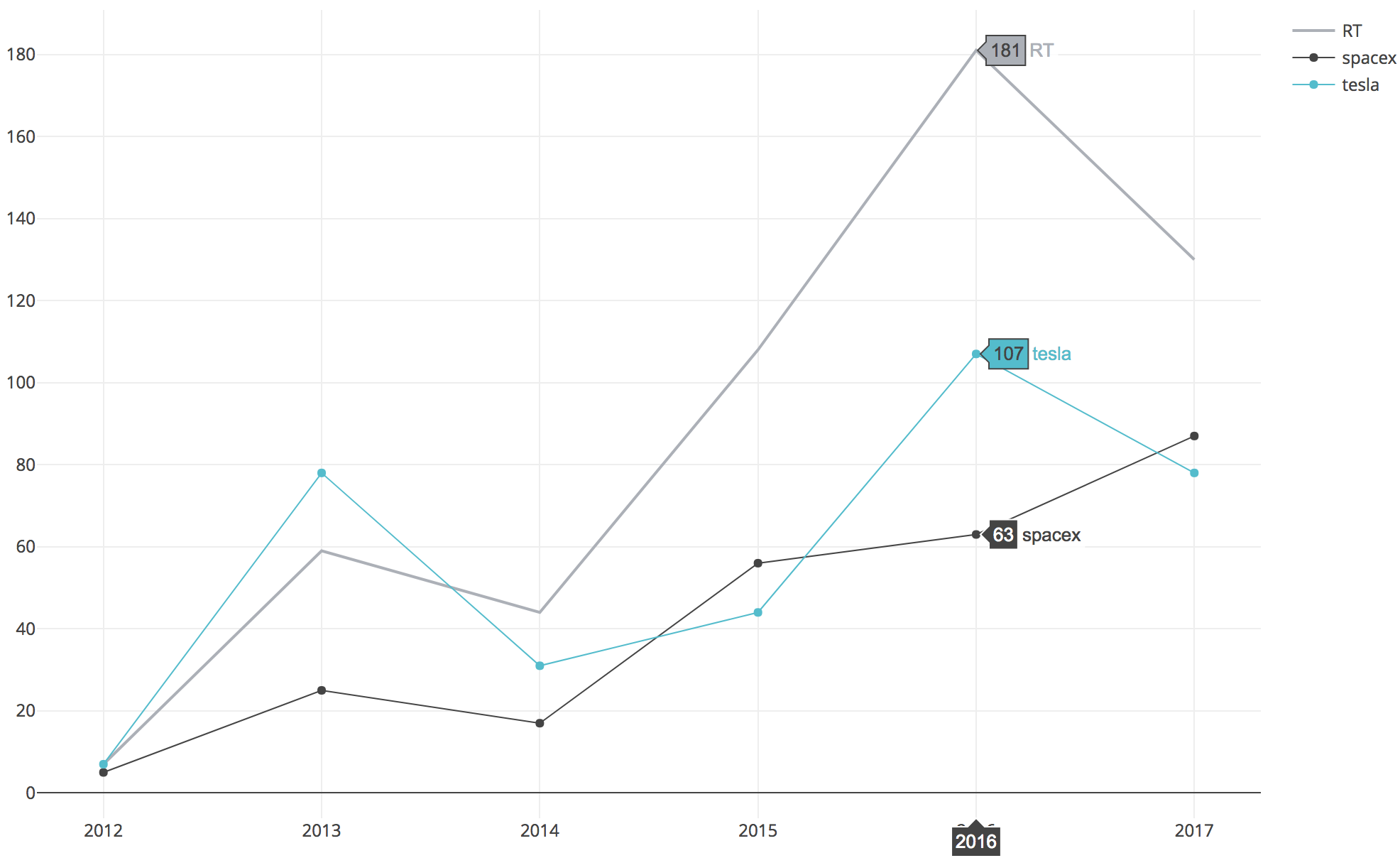
[1] https://www.kaggle.com/datasets

[2] https://finance.yahoo.com/

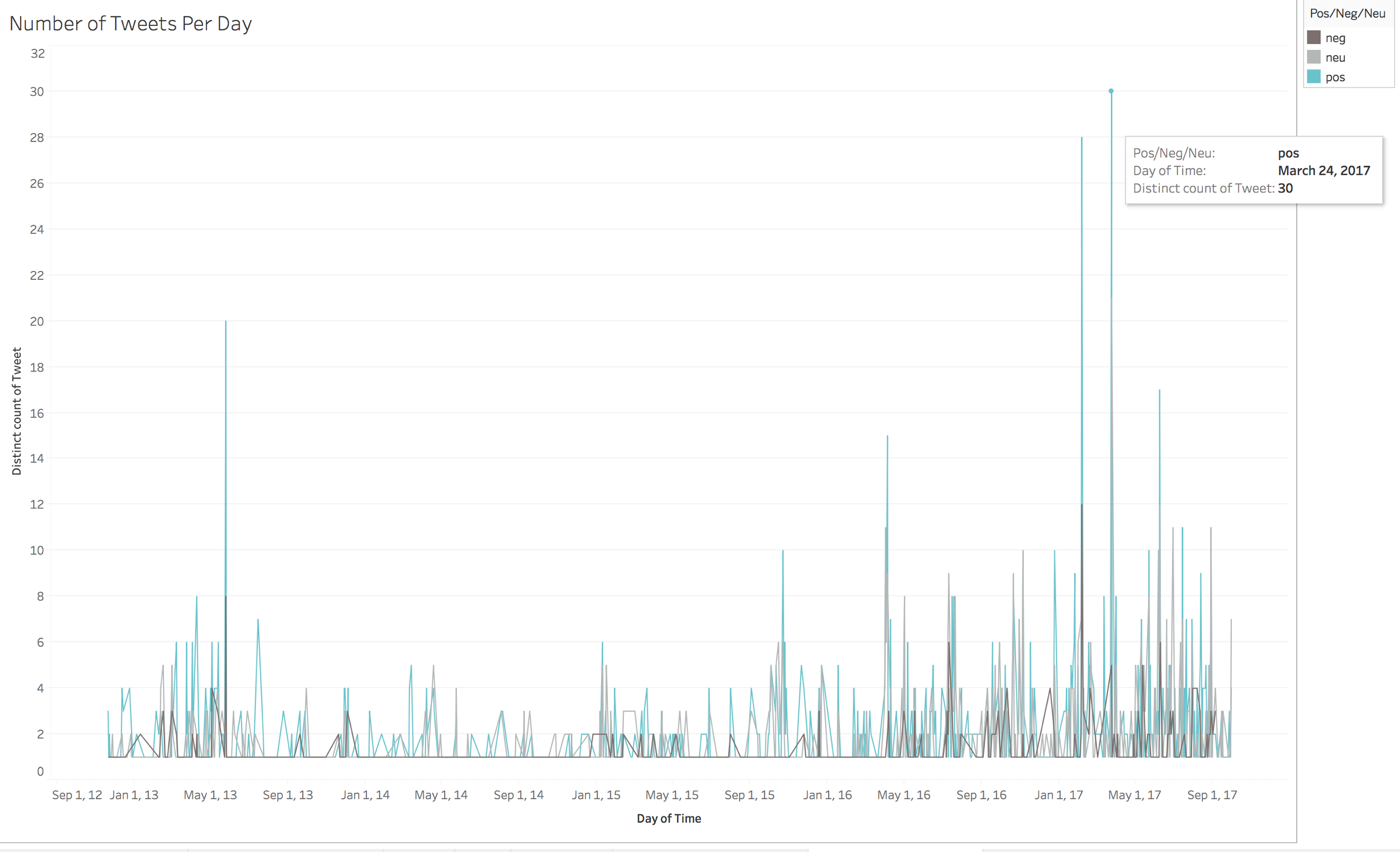
[3] http://Stocktwits.com/home

**Appendix.**

Pic. 1

****

Pic. 2

****