CAPSTONE MILESTONE REPORT

Date: 20 August 2016

Author: Aakant Taurani

I. Introduction

A. What is the problem you want to solve

The problem is how to identify and measure which companies are creating sustained value for their consumers so that better investment decisions can be made.

B. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

The client is any person looking to invest funds in a particular, publicly-listed company. They care about this problem because if they do not accurately identify the companies that are creating value for their consumers, they might invest in companies that do not really create value for consumers and might eventually end up losing their investments. Based on the proposed analysis, the clients can make better decisions whether or not to invest a particular company.

II. Data Exploration

A. Datset description:

The dataset is a collection of the most recent 3200 tweets for each of 19 reputable news networks from the Twitter website. The most important fields contained in each tweet are:

1. created_at – This is the creation date of the tweet
2. favorite_count – This is the number of "favorites" a tweet has received
3. retweet_count – This is the number of times this particular tweet has been retweeted
4. is_quote_status – This boolean field explains whether a tweet is a quote tweet or not. A quote tweet is a tweet in which the user has copied another user's tweet and added a few words to the tweet.
5. text – The text of the tweet
6. user – The username of the account that posted the tweet

B. Limitations of the dataset

This dataset does not contain the sentiment of each tweet. Hence, a sentiment analysis will have to be run on each tweet to gain sentiment information. Also, in order to predict the stock market, this dataset will be used along with the dataset about stock information from Yahoo Finance.

III Inferences

A. Findings

1) There is a statistically significant difference between number of favorites for quoted tweets and the number of tweets that are not quoted tweets. If a tweet is quote, then the favorite count is very low (average of 33 favorites). If a tweet is not a quote, then the favorite count is higher (average of 90 favorites)

2) Interestingly, if the same test is performed on each network, the p-value > 0.05 for all of them. That means, individually, there is not a significant statistical difference in the number of retweets between a quote tweet and a regular. However, overall, as seen earlier, there is a significant statistical difference overall. A conclusion from this observation is that twitter users do retweet fewer times if a tweet is a quote, irrespective of news source.

3) There are atleast 2 news sources for which there is a significant relationship between favorites on a tweet and whether the tweet is a quote or not. As seen in the table, for NPR and The Real News, there is a significant drop of "favorites" if their tweet is a quote tweet vs if it is not a quote tweet.

4) There is statistically significant evidence that there is a relationship between a posts that were posted before 10am and posts after 10am. For posts before 10am, the number of "favorites" posts receive is higher than posts after 10am.

B. Approach review

Since hypothesis testing is a great way to understand whether an observed trend is significant or not, I plan to use more hypothesis testing for the capstone project, especially the 2-sample t-tests.