**Analysis on the Effect of Targeting Hillary in Donald Trump’s Election Tweets**

**Author: Dieter K. Haage**

**ABSTRACT:**

1. *Introductory Statement*: This study covers Donald Trump’s tweet history throughout the U.S 2016 election process, looking at the important issue of what effect targeting his opposing candidates had on the popularity of his tweets. This report addresses the effectiveness of aggressive campaign tactics in his presidential election.
2. *Purpose*: The objectives of this report are to identify if targeting Hillary Clinton in tweets contributed to overall tweet popularity throughout Donald Trump’s campaign, and if so to what extent.
3. *Methodical Approach*: The data used in this report includes the tweet history for Donald Trump’s Twitter account over the election period, listing the text within each tweet. Methods used in this report include Data Gathering, Cleaning, Selection/Sampling, Type Conversion, Variable Transformation, Exploratory Visualization, and Group Based Summarization & Visualization. Tweet data ‘Favourite’ and ‘Retweet’ were used to measure popularity, and correlated to tweets including ‘Hillary’.
4. *Findings or Achievements*: Findings for this study contain; the effect that targeting Hillary had on Trump’s tweet popularity. The frequency of specific word usage throughout Trump’s tweets.
5. *Conclusions and Implications*: Targeting Hillary in tweets did not, on average, increase the popularity of the tweets posted by Donald Trump over his election campaign. Findings of this report deliver insights into certain social media campaign strategies, and their effectiveness in political elections.

**INTRODUCTION:**

Over the 2016 U.S election period, Donald Trump campaigned for his candidacy for President of the United States (P.O.T.U.S.). His social media campaign strategies caused controversy, and may have been justified in the sense that by targeting his opponents on Twitter, he gained popularity and exposure which contributed to the success of his campaign and his election as P.O.T.U.S.

The motivation for this report is to gain insights into whether or not the strategy contributed to Donald Trump’s campaign success.

The objectives of this report are to identify the effect that targeting Hillary Clinton in tweets had, and the contribution to overall tweet popularity throughout Donald Trump’s campaign, and to what extent.

The position of this report is; the effect of targeting Hillary Clinton, increased the popularity of Donald Trump’s tweets on average.

The outcomes of this report should include insightful and conclusive evidence that allow us to determine the effectiveness of Donald Trump’s campaign strategies, and whether or not it has a place in the future of presidential election campaigns.

**DATA:**

The data used in this report includes the tweet history for Donald Trump’s Twitter account over the 2016 U.S presidential election period.

1. *Data Source*: The source of the data (Trump-Tweets!.csv, 2018) is Donald Trump’s Twitter account (@realDonaldTrump), covering tweet history from 16/07/15 to 11/11/16.
2. *Data Collection*: The data was originally collected from an observational study on the website Kaggle (Trump-Tweets!.csv, 2018) as a .csv file, where the original author is not known. The separate study had initially only focused on Trump’s tweet rate and frequency.
3. *Sample Size*: the size of the data set Donald-Tweets!.csv is 1.62MB with 7375 Observations (Tweets).
4. *Number and Types of Variables*: The original data (Donald-Tweets!.csv, 2018) includes 12 different variables, they are listed in Table 1. below along with the variable type:

Table 1.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type** |
| Date | Tweet Date | Numeric Date Time |
| Time | Time of Tweet | Numeric Date Time |
| Tweet\_Text | Tweet Text | Nominal |
| Type | Link/Text | Categorical Nominal (Binary) |
| Media\_Type | Device Used | Categorical Nominal |
| Hashtags | Hashtags Included | Nominal |
| Tweet\_Id | Tweet ID | Categorical Nominal |
| Tweet\_Url | Location of Tweet | Nominal |
| twt\_favourites\_IS\_THIS\_LIKE\_QUESTION\_MARK | Number of ‘Favorited’ on tweet | Discrete Numeric |
| Retweets | Number of ‘Retweets’ on tweet | Discrete Numeric |
| X | Missing Values | NA |
| X.1 | Missing Values | NA |

1. *Known interventions/pre-processing preceding report*: There are no known interventions or pre-processing on the data set (Donald-Tweets!.csv, 2018), other than it was selected for the original observational study after similar datasets proved to be less complete and clean.
2. *Additional Information*: The original dataset has missing values and incomplete data, and in order to view any ‘emoji’s’, a decoder may be required as well.

The Tweet\_Text variable does not contain missing values, neither does the corresponding number of ‘Retweets’ or ‘Favorited’ variables, and so for the purpose of the study, the dataset meets the necessary requirements for Data Science methods used to process and analyse.

**METHODS**: The following steps were performed to pre-process and explore the data using *RStudio.Version(3.4.3)*. The targeted key topics are numbered as follows: [1]Representation, [2]Unstructured to Structured, [3]Cleaning ,[4]Type Conversion,[5]Missing Value Imputation, [6]Gathering/Spreading,[7]Subset Selection,[8]Group Based Summarisation, [9]Variable selection/Transformation, [10]Exploratory Visualization using ggplot2.

**Step 1: Setup R-Studio Environment**, packages which will be used throughout the project; “dplyr”, “ggplot2”, “tidyr”, “stringr”, “data.table”, “tm” and “wordcloud” are installed and libraried.

**Step 2: Gather Data**, using the read.csv() function, the .csv file “Donal-Tweets!.csv” is imported including Headers and stringsAsFactors as arguments to create and name a new data frame. [6]

**Step 3: Create New Column**, using the “dplyr” String Detect function, str\_detect(), a new column is created that assigns a True/False value to each tweet that contains the text “hillary”. This gives an indication as to which tweets included Hillary Clinton as a target of the tweet. ‘True’ indicating that she was targeted in the tweet, and ‘False’ indicating that she was not targeted in the tweet. This column is used as the primary focus in producing the results. Remove Unwanted Data, columns with no values such as ‘X’ and ‘X.1’ are removed, along with other irrelevant and incomplete variables such as ‘Tweet\_Url’, ‘Tweet\_Id’, ‘Hashtags’, ‘Media\_Type’, and ‘Type’. The base R function of concatenating the columns with a ( - ) symbol preceding the column number, removes the unwanted columns.[9][3]

**Step 4: Summarise True/False Values**, using the ‘dplyr’ filtering function, filter(), a new data frame is created that shows only the Tweets for which the tweet text includes ‘hillary’. The same is done for False values. This returns two separate data frames indicating that there are a total of 656 observations for values that are True, and 6719 observations for values that are False. This provides an indication as to the total sample sizes for each of the T/F values. Using the ‘dplyr’ Rename function, rename(), the variable name of ‘twt\_favourites\_IS\_THIS\_LIKE\_QUESTION\_MARK’ is changed to “Favorited”. [8][1]

**Step 5: Transform Variables**, define the Retweet and Favorited variables using as.numeric(), and the T/F column that was created using as.factor(). Define the Date variable using as.POSIXct(). Using the na.omit() function handle missing data. Create new variable that multiplies individual Retweets by Favorited to produce a ‘Popularity’ score column. [9][3]

**Step 6: Explore and Visualise Data**, using the ‘ggplot2’ plot function, geom\_point(), create a scatter plot (Plot A.) with ‘Retweets’ as the X axis, and ‘Favorited’ as the Y axis, the title being “Tweet Popularity”, and the grouping being based on the True or False values that were previously created. Also using ‘ggplot2’ and the geom\_bar() function, create a bar plot (Plot B.) for the total popularity score for True/ False values. [8][10]

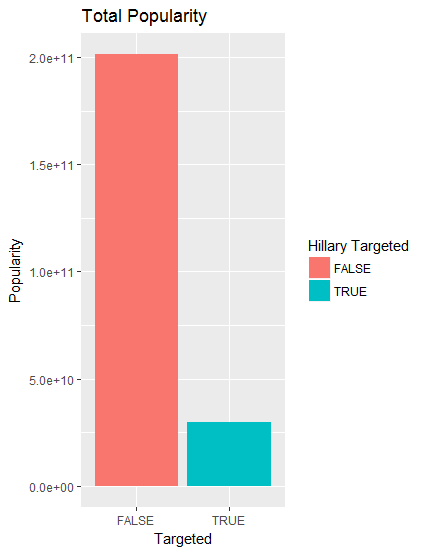
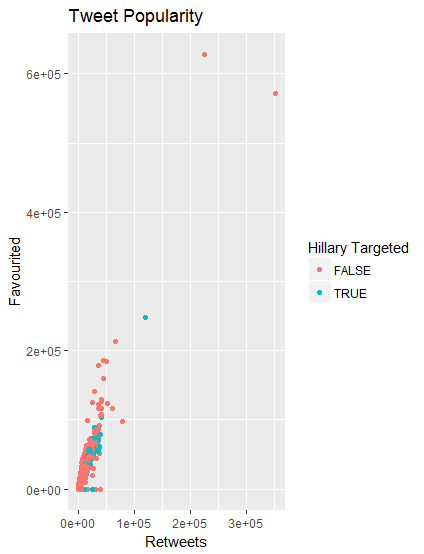
Using the ‘ggplot2’ function, geom\_bar(), create another bar plot (Plot C.), this time calculating the mean for the ‘Popularity’ variable based on T/F from the previously created variable.[9][10]

**Step 7: Text Mining**, use the ‘tm’ Paste function, paste(), to collapse all spaces in the Tweet Text column to combine all the tweets together. Set up source and corpus using functions VectorSource(), and Corpus(). Clean the data by using tm\_map to transform the text to lower case, remove punctuation, white space and stopwords (Feinerer, 2017). [2][1][3]

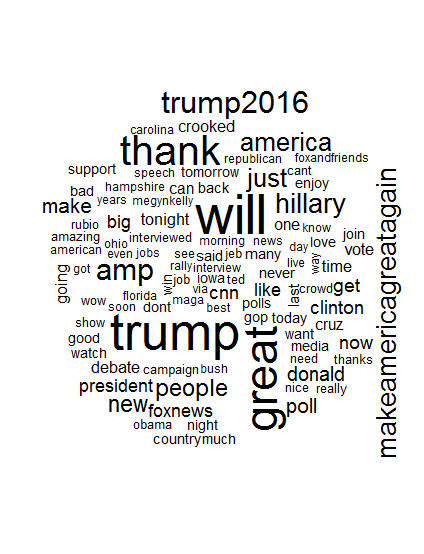
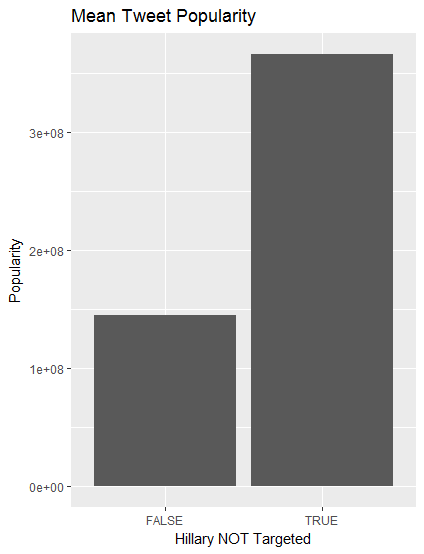
Create a document term matrix applying the DocumentTermMatrix() function to corpus, and create a new data frame that is defined using as.matrix. [2][4][8]

Find the frequency of terms by using colSums() and sort(). Visualize by using the ‘wordcloud’ functions names(), and wordcloud() for the top 100 terms and frequencies (Plot D.).[2][8][9][10]

**RESULTS & DISCUSSION**:

****(Plot A.) (Plot B.) ***Plot A****.* is a scatter plot that illustrates the popularity of Trump’s tweet depending on whether or not he targeted Hillary in the tweet. ‘True’ meaning that Hillary was targeted, and ‘False’ if she was *not*.

This outcome is useful because it demonstrates the effect that targeting Hillary had in relation to not targeting Hillary. The result aligns with the goal of understanding the effectiveness of aggressive campaign tactics. From an analysis standpoint, Plot A. shows that there are more tweets that do not target Hillary with higher popularity scores, although they are less focused. The fact that the Popularity scores are higher for ‘non-targeting’ tweets is interesting since it conflicts with the original position that the opposite would be true.

**** (Plot C.) (Plot D.) ***Plot B****.* is a bar plot that illustrates the total value of Popularity for targeted vs. not targeted. ‘True’ Meaning that Hillary was targeted, and ‘False’ meaning that Hillary was *not* targeted.

This outcome is useful because it demonstrates the sample size in terms of Popularity for each grouping (Baumer, Kaplan & Horton, 2017) .The result aligns with the goal of understanding total contribution in Popularity for Targeting vs. Not Targeted. From an analysis standpoint, there were more tweets that did not target Hillary, however it is possible to see how much focus there was on targeting Hillary in relation to the total amount of tweets. The fact that Targeting Hillary was responsible for a large portion of the total tweets is interesting since it shows to what extent the tactic was used. This supports the notion that it was in fact a strategic campaign tactic.

***Plot C****.* is a bar plot that illustrates the average popularity based on Targeted vs. Not Targeted. ‘False’ meaning Hillary was targeted, and ‘True’ meaning that she was *not* targeted.

This outcome is useful because it indicates the average amount of Popularity based on whether or not Hillary was targeted. The result aligns with the goal of understanding how effective targeting Hillary in tweets was. From an analysis standpoint, there is a higher average of Popularity for when Hillary was not targeted. The fact that there was more popularity on average when *not* targeting Hillary is interesting since this contradicts the original position of the report.

***Plot D****.* is a word cloud that illustrates the frequency of terms and their usage through all of Trump’s tweets. The highest frequencies appearing as the larger terms in size.

The frequency of the term “Hillary” was 532, and the frequency of “Trump” is 1188. This is useful in determining the level to which trump focused on targeting Hillary. The fact that he targeted Hillary directly 532 times and mentioned himself nearly twice as much is interesting since we can determine the potential value gained for both the terms. This aligns with the goal of understanding the tactic’s effectiveness.

**CONCLUSION:**

The original objectives of this study were to understand the effectiveness of aggressive campaign tactics through Donald Trump’s tweets, and to what extent. The original position of the study was; that by targeting Hillary Clinton in tweets, there was a positive effect in popularity. The results of the study contrast the original position on the following aspects:

1. The popularity scores for targeting Hillary were lower than when Hillary was *not* targeted.
2. There were more Retweets, and more Favorited when Hillary was *not* directly targeted.

In summary; it can be concluded that, based on the findings, although targeting Hillary Clinton played a large role in Donald Trump’s campaign tactics, it had less impact than tweets that did not target Hillary Clinton directly.

As a recommendation for future political election campaigns, it would be more effective in terms of increasing tweet popularity by focusing more on higher yield tweets as opposed to dedicating a large portion of tweets to targeting the opposition directly. Unavoidable limitations for this investigation include; the overall sentiment and rhetoric of Donald Trump’s tweets, and the correlation between major events over time for the posted tweets.

**REFERENCES:**

Larson, L (2017, April 16). *Donald Tweets.* Retrieved from https://www.kaggle.com/kingburrito666/better-donald-trump-tweets/data

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Baumer, S.B., Kaplan, T.D., Horton, J.N., (2017). Modern Data Science with R. *Texts in Statistical Science,* (12-13). Boca Raton, FL: CRC Press.

Feinerer, I. (2017). Text Mining in R, *Introduction to the tm Package*, retrieved from https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf