MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

Twitter Data Mining and Predictive Modeling in R

Eliana Espinosa (E-mail: eespinosa@stu.edu) [a],
Reinaldo Sanchez-Arias (E-mail: rsanchez-arias@stu.edu) [a]
[a] *School of Science, St. Thomas University, Miami Gardens, FL, USA.*

Graphical Abstract

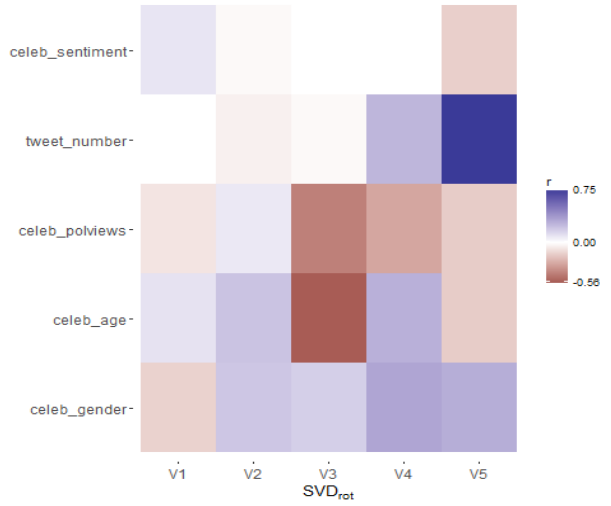

Fig. 2: Correlation of Singular Value Decomposition (SVD) dimensions and psych demographic traits

Keywords: *social media, sentiment analysis, Twitter data mining.*

Abstract.

The open source statistical programming language R can be used to gather information from the social media platform Twitter, to collect tweets from various news sources, celebrities, political figures, official colleges accounts, among others. Information such as screen names, number of tweets, number of followers, list of friends, and locations can be collected using the twitteR package in combination with the Twitter application programming interface (Twitter API). With such data, one can perform text mining by counting the word frequency in news sources' tweets, creating data visualizations to represent frequency of words, and conduct a sentiment analysis to understand and measure the impact of certain topics and opinions expressed in this social media venue. This project explores the various ways that Twitter can be used to gather information on certain topics and how this data could be used to help predict some of the behaviors and characteristics on how people communicate through this social media outlet.

Introduction

We used similar methods presented by M. Kosinski et. al [1] in their work on mining big data (from Facebook users) to predict real-life outcomes, performing a data analysis of Twitter data using the open source statistical programming language R [2]. Some of the mathematical techniques used in this work include Singular Value Decomposition (SVD), and Logistic Regression models. A training matrix was generated gathering different data from celebrities including the number of people they follow, their age, gender, political party, and number of tweets. The data was retrieved using the Twitter API and the twitteR package [3] during July 2017.

Results and Discussion

We created a heatmap showing the correlation between the SVD dimensions and the psychological demographic traits of each celebrity in the dataset we created. We focused on $k = 5$ SVD dimensions. In Fig. 2 we can notice that the SVD dimension V5 correlates positively with the number of tweets and gender while it has a negative correlation with the sentiment, political views, and age.

	Name	Gender	Age	Pol. View	Tweet Num.
[1,]	"Adam Sandler"	"M"	"50"	"R"	"180"
[2,]	"Ben Affleck"	"M"	"44"	"D"	"395"
[3,]	"Beyonce"	"F"	"35"	"D"	"10"
[4,]	"Blake Lively"	"F"	"29"	"D"	"25"
[5,]	"Chris Hemsworth"	"M"	"33"	"R"	"170"
[6,]	"Christina Aguilera"	"F"	"36"	"D"	"979"
[7,]	"Cristiano Ronaldo"	"M"	"32"	"D"	"2916"
[8,]	"Hugh Jackman"	"M"	"48"	"D"	"2942"
[9,]	"Jason Mraz"	"M"	"40"	"D"	"3460"
[10,]	"Karim Benzema"	"M"	"29"	"D"	"1156"
[11,]	"Kourtney Kardashian"	"F"	"38"	"D"	"12200"
[12,]	"Liam Hemsworth"	"M"	"27"	"D"	"149"
[13,]	"Luis Suarez"	"M"	"30"	"R"	"838"
[14,]	"Madonna"	"F"	"58"	"D"	"2452"
[15,]	"Mariah Carey"	"F"	"47"	"D"	"7365"
[16,]	"Mark Wahlberg"	"M"	"46"	"D"	"1030"
[17,]	"Michelle Obama"	"F"	"53"	"D"	"798"
[18,]	"Sandra Oh"	"F"	"45"	"D"	"535"
[19,]	"Sofia Vergara"	"F"	"44"	"D"	"6819"
[20,]	"Sylvester Stallone"	"M"	"70"	"R"	"1251"
[21,]	"Tom Hanks"	"M"	"60"	"D"	"799"
[22,]	"Victoria Beckham"	"F"	"43"	"R"	"3156"
[23,]	"Zac Efron"	"M"	"29"	"D"	"1585"

Fig. 1: Sample of celebrities' data matrix and their Twitter accounts (July 2017)

Programming was essential in the development of this project. R is an open source language widely used in the data science community, with focus on statistical data analysis, data visualization and machine learning methods. During this project, tools from the tidyverse package [4] were used for data wrangling and data visualization with the help of RStudio, an open source integrated development environment (IDE) for R. Sentiment analysis can be thought of as the exercise of taking a sentence, paragraph, document, or any piece of natural language, and determining whether that text's emotional tone is *positive*, *negative* or *neutral*. Using the Twitter API, we collected information on the Twitter accounts of celebrities following less than 100 accounts (as of July 2017), and performed sentiment analysis on the messages they posted on this social media outlet. Fig. 3 shows the distribution of sentiment scores from different accounts.

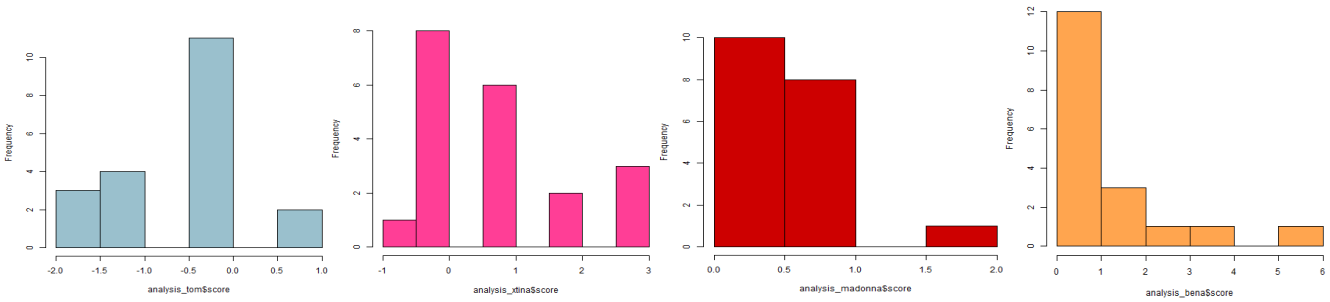

Fig. 3: Sentiment score for tweets by Tom Hanks, Christina Aguilera, Madonna, and Ben Affleck (as of July 2017)

Using the twitteR package in R, we gathered information of users who follow the official Twitter account of St Thomas University (https://twitter.com/StThomasUniv). Using the geolocation information associated to each user, longitude and latitude coordinates can be easily mapped, to have a measurement of the diversity of followers of any account. In Fig. 4 we show the locations of the @StThomasUniv followers as an example, both in North America and Worldwide.

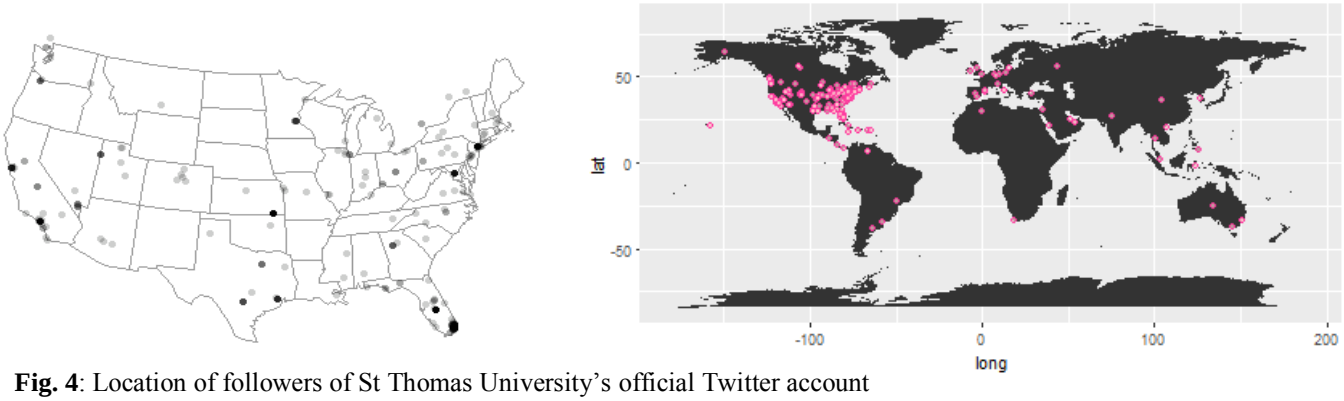

Fig. 4: Location of followers of St Thomas University's official Twitter account

Conclusions

The conjunction of R for data analysis and different APIs makes for a powerful tool for data mining and visualization of social media outlets data like Twitter.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

Authors want to thank St. Thomas University facilities for completing this work during the SRI 2017. This project was supported, in part, by U.S. Department of Education grant award P03C1160161 (STEM SPACE), P031c160143 (STEM EngInE), P120A160036 (STEM ISLE), 1161177 (STEP Up), P120A140012 (SPARC).

References

1. Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. Psychological Methods, 21(4), 493-506.
2. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org/)
3. Jeff Gentry (2015). twitteR: R Based Twitter Client. R package version 1.1.9. (https://CRAN.R-project.org/package=twitteR)
4. Hadley Wickham (2017). tidyverse: Easily Install and Load 'Tidyverse' Packages. R package version 1.1.1. (https://CRAN.R-project.org/package=tidyverse)