A Framework for Association Rule Learning with Social Media Networks

Ryan Kruse, Tharindu Lokukatagoda, Suboh Alkhushayni

Department of Computer Information Science, Minnesota State University, Mankato

E-mail: ryan.kruse@mnsu.edu

E-mail: tharindu.lokukatagoda@mnsu.edu

E-mail: suboh.alkhushayni@mnsu.edu

November 2020

Abstract. We present an application of association rule learning to analyze Twitter account follow patterns. In doing so, we develop a basic framework and tutorial for future researchers to build on, which takes advantage of the Twitter API. To demonstrate the method, we take samples of Twitter accounts following Joe Biden and Donald Trump. For each account in our sample population, we pull the account's 100 most recently followed accounts. This data is cleaned and formatted for use with Python's **apyori** package, which uses the well-known apriori algorithm to learn association rules for a given dataset. This work has two objectives: (1) demonstrate the application association rule learning to social media networks and (2) perform exploratory analysis on the resulting association rules. We successfully demonstrate association rule learning in a Jupyter-notebook environment with Python. The resulting association rules indicate some interesting similarities and differences in the networks of Biden's and Trump's Twitter followers.

Keywords: association rule learning, Twitter, social media, market basket analysis, data mining

1. Introduction

Twitter is gold mine of data where almost every user's tweets are publicly available. The social media platform gives individuals—and notable public figures—the opportunity to widely share their unfiltered thoughts. Users share Tweets with varying levels of frequency; some may Tweet several times a day, while others may only Tweet on rare occasion. The United States of America's two presidential candidates, sitting President Donald Trump and former Vice President Joe Biden, both use Twitter to regularly appeal to the public in a way that was not possible until recently. Each candidate has millions of followers, most of which follow hundreds or even thousands more accounts. This platform provides a means to analyze rule associations between accounts

by looking at the followers of each candidate.

Association rule learning is most used with transactional data—market basket analysis. In market basket analysis, grocery store purchases might be analyzed to learn the association rules between items. Recently, it has been applied to text data, including Twitter. One such application, presented by Professor Ami Gates of Georgetown University [1] at the 2019 Washington DC R Conference, learns association rules for Tweets containing specific words, such as "chocolate." Our work applies a similar method to new data in an innovative way.

In context of grocery shopping, association rule learning might be used to identify which items are most and least likely to be bought with bread at Walmart. One might discover that when peanut butter is bought, bread is also frequently bought, but when bread is bought, peanut butter might not be as frequently bought (people who buy peanut butter usually buy bread but people who buy bread don't necessarily want peanut butter with it). In our context, we can think of "following Donald Trump on Twitter" as "buying bread at Walmart," "following Andrew Yang on Twitter" as "buying peanut butter at Walmart," and "following Elon Musk on Twitter" as "buying jelly at Walmart." A "transaction" is a Twitter account's following list. Just like in market basket analysis when we want to analyze all transactions where bread is purchased, we can analyze all Twitter accounts that follow Donald Trump. Then, asking "What is the relationship between Andrew Yang and Elon Musk on Twitter among Trump's Twitter followers" is like asking "When bread is purchased, what is the relationship between peanut butter and jelly?"

Many people have recently been concerned with social media acting as "echo chambers" where users only see content from like-minded users. This is believed to contribute to increased polarization in the United States. Our project will provide a lens into the inner workings of these echo chambers. Additionally, our project will provide a foundation for future applications.

1.1. Objectives and Research Questions

The objectives and research questions of this study are:

- Demonstrate the application association rule learning to social media networks.
- Perform exploratory analysis on the resulting association rules.
  - How do the candidates' networks relate in terms of shape?
  - What are the notable account similarities and differences in the candidates' networks?

2. Related Work

The literature on Tweet mining and analysis is extraordinarily rich, much of which revolves around sentiment analysis. For example, Nesi et al. [4] analyzed Tweets in an effort to determine how likely they are to get ReTweeted. Other papers attempt to predict the political alignment of Twitter users [2][3]. Perhaps most related is Finding Influential Users in Social Media Using Association Rule Learning [5]. In this paper, Erlandsson et al. apply association rule learning to Facebook data in an effort

to identify influential users and interesting posts. Also closely related is a 2013 paper by Cagliero and Fiori [6], who apply association rule learning to Twitter posts to understand content and context of Tweets.

Our work is fundamentally different from each of these. Instead of analyzing the content of social media posts, we are interested in social media networks. Although similar to Erlandsson et al., we are working with Twitter data, which is quite different from Facebook data. Additionally, we are interested in comparing account association networks with other account association networks, instead of identifying influential accounts.

3. Methods

First, we retrieved the data using the Twitter API. To do this, we needed to obtain the proper authorization credentials. The data required were the usernames of the Twitter accounts followed by the Twitter followers of Donald Trump and Joe Biden. Because each candidate had millions of Twitter followers—each of which may follow hundreds or thousands of other accounts—and the Twitter API's strict rate limit, we decided to take samples instead of the entire population. We pulled data for a total of 227 accounts. Of these, 130 were from Joe Biden's followers and 97 were from Donald Trump's followers. The data was stored using one-hot encoding in a **pandas** dataframe. The code is available on GitHub for this step and all others [8].

Next, we organized the data to conform to Python's **apyori** package. This required restructuring the data into a list of lists, where each nested list represented one account's following list.

Then, we executed the apriori processing. In doing so, we set the necessary parameters, including **min_lift**, which filtered our results to those with greater lift than given value. This gave us the association rules for followers of each candidate.

Finally, we analyzed the learned association rules, looking for notable similarities and differences between each candidate.

3.1. Association rule learning

Association rule learning is a method for discovering interesting relations between variables in large databases. It is most commonly known for its applications in transactional data (e.g., market basket analysis). In such a use case, association rules between items such as peanut butter, jelly, and eggs are learned to better understand customer preferences and habits. The grocery store can then act on this new knowledge to better market to their customers.

Association rules are primarily composed of three metrics—**support**, **confidence**, and **lift**—which define the relationship between some set of **antecendents** and **consequents**. This section's definitions come from Tan et al [10].

Definition 1. ***Support*** *determines how often a rule is applicable to a given data set. A rule that has very low support may occur simply by chance. Support for given rule* $X \to Y$ *is given by*

$$s(X \to Y) = \frac{\sigma(X \cup Y)}{N},$$

where X *and* Y *are the antecedent and consequent, respectively;* N *is the total number*

of transactions in the dataset; and $\sigma(W \cup Z)$ is the count of transactions in the dataset containing items W and Z.

Definition 2. ***Confidence*** *measures the reliability of the inference made by the rule. For a given rule $X \to Y$, the higher the confidence, the more likely it is for Y to be present in transactions that contain X. Confidence also provides an estimate of the conditional probability of Y given X. Confidence is given by*

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

Our work focuses on support and confidence, each of which have limitations. Typically, a minimum support threshold is set to only include rules with support greater than the threshold. The appropriate support threshold may vary from study to study. Setting the threshold too high may result in interesting patterns being lost, while setting it too low may result in an abundance of uninteresting patterns being analyzed. We use a support threshold of 0.05 in our analysis.

While not a focus of this study, we would be remiss to not mention lift, as anyone who uses this framework should be aware of it.

Definition 3. ***Lift*** *computes the ratio between the rule's confidence and the support of the itemset in the rule consequent. Lift is given by*

$$Lift = \frac{c(X \to Y)}{s(Y)}.$$

Depending on the analysis, lift may be useful as it solves the problem of misleading high-confidence rules, which are enabled by confidence not taking support into account. Suppose we have a high-confidence (say, 0.75) rule $X \to Y$, which might suggest X indicates a strong likelihood of the presence of Y. However, it is entirely possible that the support of Y is 0.90, thus, greater than the confidence. In this scenario, the following are true:

- a random transaction from the entire dataset would have 0.90 probability of containing Y;
- a random transaction from the entries in the dataset containing X would have a 0.75 probability of containing Y.

In this situation, the presence of X actually *decreases* the likelihood of Y, despite the high confidence of 0.75. Lift captures this relationship, giving a value of

$$\frac{0.75}{0.90} \approx 0.83 < 1,$$

indicating Y appears less with X than without X.

3.1.1. Apriori Algorithm

Originally proposed by Agrawal and Srikant [7], the apriori algorithm is commonly viewed as the simplest algorithm for association rule learning, which uses a breadth-first search strategy. Other common algorithms include Frequent Pattern Growth (bottoms-up search) and ECLAT (depth-first search). While these algorithms have greater ability to scale, apriori is appropriate for our objectives.

Specifically, we use the **apyori** package in Python, available on GitHub [9]. This package makes the apriori algorithm accessible in Python.

4. Results

For Biden's network of followers, we found thirty rules with one item in the antecedent and seven rules with two items in the antecedent. For Trump's network of followers, we found twenty-four rules with one item in the antecedent and eight rules with two items in the antecedent.

Here is an example of three support rules for Biden's network.

rule	support
AOC → JoeBiden	0.146
BarackObama → JoeBiden	0.208
AOC, BarackObama → JoeBiden	0.062

The above rules relate to Alexandria Ocasio-Cortez and Barack Obama. In the first rule, the support of 0.146 indicates that about 14.6% of the accounts in our Biden dataset follow AOC. Similarly, the second rule indicates that about 20.8% of the accounts follow BarackObama. The third rule indicates that about 6.2% of the accounts follow both AOC and BarackObama.

Here is an example of confidence and lift rules for Biden's network.

rule	confidence	lift
AOC → BarackObama	0.421	2.027
BarackObama → AOC	0.296	2.027

In the first rule, the confidence of 0.421 indicates that about 42.1% of the accounts in our Biden dataset who follow AOC also follow BarackObama. Similarly, the second rule's confidence indicates that about 29.6% of the accounts that follow BarackObama also follow AOC. The lift indicates that an account in our Biden dataset that follows BarackObama is about 2.027 times more likely to follow AOC than an account in our dataset that does not follow BarackObama (and vise versa).

Our analysis resulted in many of these association rules. We are working to present the information in an appealing way. One visualization commonly used for association rule learning can be seen below: a scatter plot of confidence vs support. We do not think this is very insightful, so we will likely replace these graphs in our next draft.

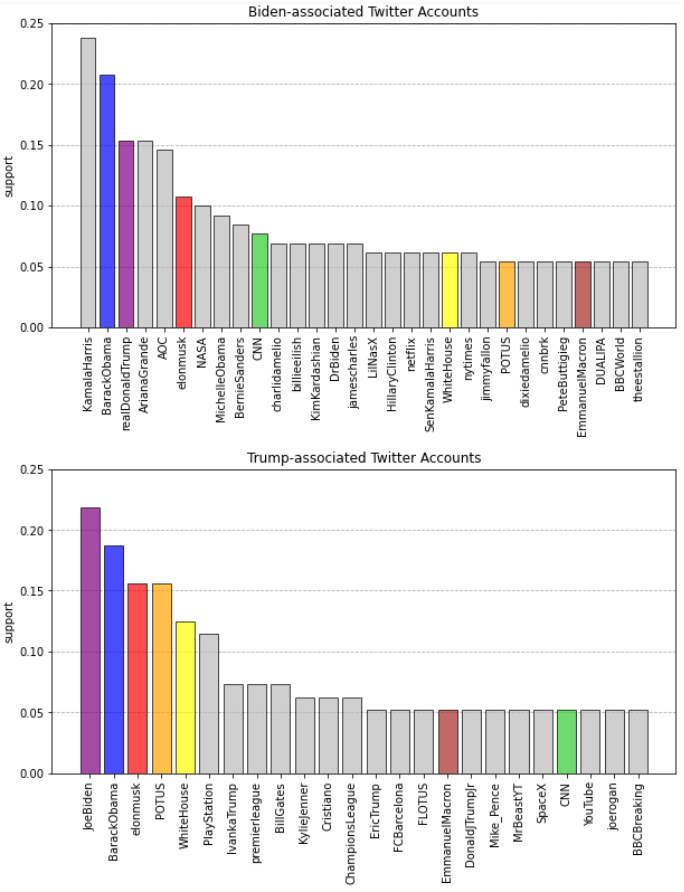

Figure 1: Most frequently appearing accounts for each candidate.

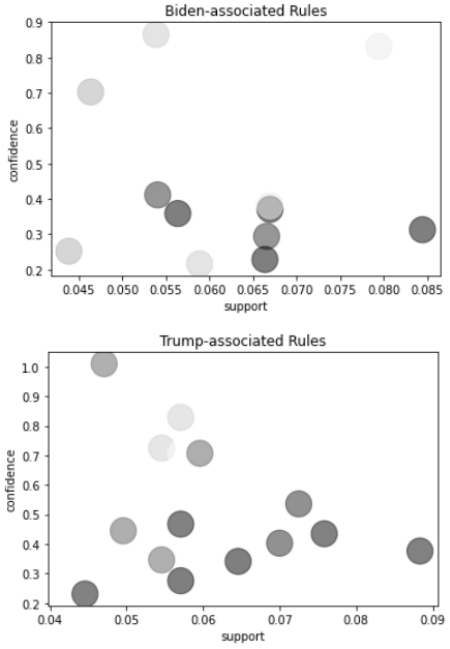

Figure 2: Plots of support, confidence, and lift, where the darker shade of gray represents a greater lift.

Finally, here we can see the relationship between support, confidence, and lift for each person's network. If there were significant differences in the overall shapes of each network of rules, they would be apparent here. There are no striking differences for these plots.

5. Discussion

Before conducting the analysis, we expected to see some striking differences between the two networks of followers. Although there were some interesting differences, we were more struck by the similarities. For example, CNN's Twitter account appeared in both Biden and Trump's association rules, with supports of .077 and .052, respectively. Additionally, each candidate appeared in each other's network at a somewhat similar rate; Biden's support among the Trump dataset was .22, and Trump's support among the Biden dataset was .15.

While the similarities were more prevalent, we did find one notable difference. Trump's dataset included a surprisingly high support (.11) for PlayStation's Twitter account. This, among the presence of other surprise accounts related to international sports, might indicate the presence of "bot" accounts—accounts attempting to act like humans when they are actually just automated.

6. Limitations and Considerations

The Twitter API presented several challenges. Most notably, we had to be careful to avoid surpassing the request rate limit. Additionally, we could only pull the most recent followers of each candidate.

We should note that at the time of our analysis, Trump had about 88 million followers while Biden had about 9 million. Trump was the President of the United States at the time, so it is no surprise that he had many more followers.

7. Conclusion

In this work, we demonstrated the application of association rule learning to Twitter account following networks. In doing so, we developed a basic framework for future researchers to follow and build on. We also showed the potential of this application by analyzing Joe Biden's and Donald Trump's network of Twitter followers. The analysis was intended to be preliminary and exploratory, but it still resulted in some inter-

esting insights that highlighted similarities and differences among the followers of each candidate.

This was a novel application of association rule mining. The technique is primarily used in market basket analysis (if you buy X, you may be interested in Y) or recommendation systems (if you watch movie X, you may be interested in movie Y). However, we recognized an opportunity to apply the technique to a new field.

There are many opportunities for future work. First, one could work with a larger dataset. This requires either (a) time or (b) money, as the Twitter API has strict rate limits for the free version. However, bigger datasets will reveal much more powerful association rules. In fact, this is where the true potential of our work lies. Applying association rule learning to a large dataset of Twitter follower relationships will reveal hidden relationships and other insights. Second, one could, of course, apply our method to other "target" Twitter accounts. We focused on Joe Biden and Donald Trump, but any number of Twitter accounts could be analyzed in a similar way. Third, this work shows that association rule learning has applications outside of the traditional use cases. With the growing availability of data, we suspect there to be many use cases in fields such as politics, education, public health, and more. In fact, with the right data, association rule learning could even be be helpful in learning how pandemics travel.

8. References

[1] Lander Analytics. (Mar 26, 2019). Association Rule Mining With Tweets: Thinking Outside the Basket [Video]. YouTube. https://www.youtube.com/watch?v=eOOhn9CX2qU

[2] Conover, Michael D. et al. "Predicting the Political Alignment of Twitter Users." 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (2011): 192-199.

[3] Makazhanov, A., Rafiei, D. & Waqar, M. Predicting political preference of Twitter users. Soc. Netw. Anal. Min. 4, 193 (2014). https://doi.org/10.1007/s13278-014-0193-5

[4] Nesi, P., Pantaleo, G., Paoli, I. et al. Assessing the reTweet proneness of tweets: predictive models for retweeting. Multimed Tools Appl 77, 26371–26396 (2018). https://doi-org.ezproxy.mnsu.edu/10.1007/s11042-018-5865-0

[5] Erlandsson, Fredrik & Bródka, Piotr & Borg, Anton & Johnson, Henric. (2016). Finding Influential Users in Social Media Using Association Rule Learning. Entropy. 18. 164. 10.3390/e18050164.

[6] Luca Cagliero, Alessandro Fiori. Discovering generalized association rules from Twitter. 2013 July. IOS Press. Volume 17, Issue 106.

[7] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.

[8] R. Kruse, November 2020. [Online]. Available: `https://github.com/kruser1/twitter-apyori`

[9] Y. Mochizuki, November 2019. [Online]. Accessed on: November 2020. Available: `https://github.com/ymoch/apyori`

[10] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2019. Pearson. Introduction to Data Mining, 2nd Edition. Chapter Six: Association Analysis: Basic Concepts and Algorithms.