Misinformation on Twitter

G02
Alex Day, Megan Fowler, Xiaohong Zhang
CPCS 4030/6030, Fall 2020

Project Report

Introduction

The primary goal of this project is to explore the differences in information propagation on Twitter as it applies to misinformation spread. Information can be broadly separated into the categories of true and false statements. We found that networks of agents spreading these two classes of statements differ in both how the information diffuses over time and how many primary actors spread the information. By visualizing the differences we presented a novel and intuitive method of visual classification of the networks into these categories. Our website first guides the user explicitly through these differences, ending with a network visualization that evolves over time so that the user can clearly see the differences in both network structure and primary actors/clusters. The website finishes with a five question quiz to convince the user that these graphs can be classified based on these aspects alone rather than requiring a more complex analysis.

Dataset

The dataset we used for the project consisted of a set of networks constructed by tracking information propagation centered on one source tweet. The Twitter15 dataset has been used in rumor detection [7] and contains 1,490 source tweets. For each source tweet, there is a tree structure built based on the spreading of this source tweet, where the source tweet is treated as the parent node, and a retweet is a child node. Each node has three data fields: 'uid', 'tweet ID', and 'post time delay (in minutes)'. There is a directed edge between the parent and child node, indicating the interaction between the two tweets such as retweet, reply, etc. Each source tweet and resulting network was assigned a label regarding the content of the tweet: True, False, Unverified, and Non-Rumor.

Design solution

The first visualization we chose was a bar chart with the average time it took for a tweet to be retweeted for the first time in the two different categories we were focusing on (false and non-rumor). We chose to open with this chart because it displayed important information about the temporal aspect of the network while also being easy to understand. This visual depends on the categorical attribute of type of tweet and impacts the coloring of the bar. This is important because this color scheme is carried throughout the project to tie each visual together for the user.

The bar chart showed us that there was a large discrepancy between the average retweet time for tweets labeled as false, compared to those labeled as non-rumor. From there, we wanted to explore potential causes of this discrepancy. We decided to create what we call a Packed Pie Chart. This visual contains a pie chart for the types of tweets shared by each individual user. The size of the pie chart correlates with the number of tweets shared. So the larger the pie chart, the more active the user was in retweeting. For the purpose of exploring the finding in the bar chart, we chose to only explore the sharing of false and non-rumor tweets. The resulting attributes are the individual user (categorical), retweet count (quantitative), Type of tweets (categorical). For consistency, the same colors for the bar chart were continued in this visual.

We chose to use this visual because it allowed us to look at the retweet behavior of each individual user. Did users gravitate towards one type of tweet? Could the type of tweet shared by a user be left to chance? It allows us to quickly identify users who were active in the community and the type of tweets that they shared. From this visual, we saw that there were more users that shared the non-rumors, and that often users would lean towards sharing one kind of tweet. Not saying that it exclusively happens because there were instances where users shared multiple types of tweets.

We attempted to do this in Tableau, but we found that the functionality was too limited. It resulted in a bubble chart where users who shared multiple types of tweets would have multiple bubbles. The size of each bubble representing the number of that type of tweets they shared. It resulted in a visual that may not be true to the data and story being told.

While viewing user tweeting habits is interesting, we wanted to explore how specific tweets move throughout its network. Not just where it goes, but the delay time for each retweet. To do this, we built a force directed graph. Each retweet is a node in the graph, with the source tweet being the parent node. The retweet time is represented by the link between nodes. The color of the nodes indicates the type of tweet that is being shared, with the source node being a different color so it is easily identifiable. Since a network is following a specific tweet, all nodes within the network will be the same color (except the parent node). Once again, we used the same color scheme as the previous graphs for continuity. This visual also included an animation aspect. The user is able to observe how the network builds as time progresses. They also have control over the animation so they can review the material as many times as needed, and can pause to compare the two.

This visual allows users to visualize the different patterns for different types of tweets. Since the false tweet and non-rumor tweet has different average retweet time, do they have different spreading patterns? Through the animation of the graphs, we can visualize the dynamic propagation of retweet from different types of source tweets over time. It shows the evolution of a second primary actor for the non-rumor tweet which is different from the false tweet in addition to a quicker build time for the network. This falls in line with the previous findings that non-rumors do in fact spread more quickly than false tweets. The strength of the link length is applied according to the retweet time, thus better visualization is obtained than the same strength applied to all nodes.

Since the force-directed graph displayed a clear difference between the two types of networks, we decided to include an assessment on our site where users would be tasked with identifying the type of network they are viewing. This 5 question quiz provided an interactive aspect to our site to help users engage with the material they just learned.

Literature review

The bar chart is a fundamental visual representation that is taught to students from an early age. It's a simple, straightforward method to convey information. Cleveland & Mcgill found that it is easier to interpret data that relies on position or length rather than processing the angle and curvature of a pie-chart [3]. Even though they present the same information, the bar chart is easier to interpret.

The packed pie chart is not a conventional visual which led to some difficulty finding literature. So we broke down the aspects of the chart, starting with the bubbles. A bubble plot is a scatter plot with the third dimension, size of dot [5]. Rosling utilized this type of visual in his Ted talk [9]. In his presentation, the size of the bubble represented a country's population while the x-axis was the family size and the y-axis was life expectancy. The use of the bubble chart aided the visualization because it made it easier to distinguish nations and track the changes in position of the bubbles.

Since there is not a direct relationship between the users for the second visual, the X and Y dimensions are removed. In d3, this is called a packed chart. While the Cleveland & Mcgill paper taught us the value of using length to interpret data instead of processing the curvature of angles in a pie chart, it would not be sensible to try to turn each user bubble into some form of bar chart. Therefore, the pie chart prevailed. There is evidence that pie charts are helpful when communicating information such as observed in the Van der Linden 2014 study regarding how to communicate scientific consensus on climate change [12]. They found that pie charts and descriptive text were more useful than metaphors when communicating these concepts and that appears to be due to the ability of individuals to recall the content.

Our network graph is built mainly on three forces: forceCollide, forceManyBody and forceLink. There are different algorithms behind this[4, 10]. The forceCollide uses the radius of each node to allow/disallow nodes overlapping. It is similar to the node collision to a wall, where an elastic collision or inelastic collision can be applied. The inelastic collision enables nodes to stick to each other until the force vector on the nodes has only components moving away from others. The elastic collision causes "bounce" many times, which is what we see in our network graph. However, it entails extensive computation. The forceManyBody causes all elements to attract or repel one another until it reaches a stable state. Usually repulsive forces are calculated between every pair of nodes, but attractive forces are calculated only between neighbours. The forceLink pushes the source node and the target node to be a fixed distance apart. According to different algorithms, some can move only one node at each iteration, and others can move more nodes at each iteration.

While not technically a new visual, we thought it was important to include a brief explanation for why we wanted to include a quiz that utilizes the force-directed graph. Formative assessment, such as this quiz, has been shown as a useful tool to help monitor student progress while remaining low stakes. This helps instructors to identify strengths and weaknesses so they can provide tailored assistance [11]. In today's world with the reliance upon online learning, immediate feedback has proven to be helpful to students in higher education as opposed to when feedback is delayed in the classroom setting [1]. Our quiz follows this idea by immediately providing feedback after the user submits their answer.

While our visuals convey large amounts of information to the user, the layouts are rather simple, which allows for use to accommodate users who may be colorblind. There are efforts such as seen in [6] where researchers are attempting to apply algorithms to documents in order to make them accessible to color blind viewers. Since red-green blindness affects 99% of all color blind people [2], our initial visuals were not particularly helpful. We chose to go with red and blue since those are different enough regardless of the type of blindness [8].

References

[1] Baleni, Zwelijongile Gaylard. "Online formative assessment in higher education: Its pros and cons." Electronic Journal of e-Learning 13.4 (2015): 228-236.

[2] Cravit, R. (2020, October 30). How to Use Color Blind Friendly Palettes to Make Your Charts Accessible. Retrieved December 04, 2020, from https://venngage.com/blog/color-blind-friendly-palette/

[3] Enskog, K. (2017, January 10). CLEVELAND & MCGILL – Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods (1984). Retrieved December 08, 2020, from https://creativeartsadventure.wordpress.com/2017/01/02/cleveland-mcgill-graphical-perception-theory-experimentation-and-application-to-the-development-of-graphical-methods/

[4] Fruchterman, T.M. and Reingold, E.M., 1991. Graph drawing by force-directed placement. Software: Practice and experience, 21(11), pp.1129-1164.

[5] Holtz, Y. (n.d.). Bubble plot. Retrieved December 08, 2020, from https://www.d3-graph-gallery.com/bubble.html

[6] Jefferson, Luke, and Richard Harvey. "Accommodating color blind computer users." Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility. 2006.

[7] Ma, Jing, Wei Gao, and Kam-Fai Wong. "Detect rumors in microblog posts using propagation structure via kernel learning." Association for Computational Linguistics, 2017. https://github.com/majingCUHK/Rumor_RvNN

[8] Nichols, D. (n.d.). Coloring for Colorblindness. Retrieved December 08, 2020, from https://davidmathlogic.com/colorblind

[9] Rosling, H. (n.d.). The best stats you've ever seen. Retrieved December 07, 2020, from https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen/details

[10] Tom Roth, D3 resources. https://tomroth.com.au/d3/

[11] University, C. (n.d.). Formative vs Summative Assessment - Eberly Center - Carnegie Mellon University. Retrieved December 06, 2020, from https://www.cmu.edu/teaching/assessment/basics/formative-summative.html

[12] Van der Linden, Sander L., et al. "How to communicate the scientific consensus on climate change: plain facts, pie charts or metaphors?." Climatic Change 126.1-2 (2014): 255-262.