Optimizing Outrage

Measuring Algorithmically-Induced Anger in Social Networks

Daniel Olmstead
W241.3 Hughes

Since the 2016 election, there have been a great deal of studies and articles highlighting the increasingly fractious nature of American politics. These are often distilled down to anger - the anger of the white working class towards coastal elites; the anger of long-suffering women; the anger of aggrieved white men towards immigrants; the anger of many, many people against Donald Trump. Frequently, when people discuss the cause of all this fractiousness, they point their finger at social media: Facebook, Twitter, YouTube, and other social media sites have created an "echo chamber" of "filter bubbles," they say. In their drive to increase engagement (which in turn increases advertising revenue), social networks have discovered that nothing drives clicks like outrage, and whether by design or by accident, the algorithms they have built to increase engagement have consequently led to increased outrage and division among groups. But is this true?

Core Question:
To what extent does Twitter's ranking algorithm create outrage among its users?

Why is this interesting?
Social media's potential responsibility in fomenting anger in Americans and driving them towards ideological extremes is a broad, thorny issue. But one of the most common hypotheses advanced by critics is the algorithmic design of the networks. Social networks are supported by advertising revenue, so it is in their interest to keep users engaged and returning to the site. To that end, what were initially simple chronological timelines of the network feeds evolved into algorithmically driven feeds, wherein the most "interesting" or "relevant" content to a user was promoted to the top of the feed. Typically, the algorithms have relatively simply heuristics to determine what makes content relevant or interesting - if it has engaged a lot of people who like the stuff that you like, chances are you will like it too. However, in this context "like" does not always mean "feel favorably towards"--people will retweet or respond to articles or posts that they disagree with, or that make them feel slighted, insulted or victimized. Content producers have learned that this anger drives clicks, and some have optimized their content to take advantage of it, resulting in an ecosystem of content all vying for attention at the top of users' feeds. It is an open question as to how much responsibility the ranking algorithm bears in this environment--presumably producers would be vying for clicks regardless of where they showed up in your feed. But the critics contend that the ranking algorithm creates a vicious circle, incentivizing dishonesty from critics and instilling feedback cycles of emotional release in users. Isolating the algorithm might help determine some extent of its culpability.

The ideal scenario to test this would involve a way to isolate and negate the algorithm that drives a user's feed without the user's knowledge, and then an outcome metric that did not rely on self-reported data. This design is impossible for anyone outside of Twitter, but we can get pretty close.

Recruitment
The study group will be comprised of regular users of Twitter--ideally, people who meet the following criteria:

- Identify as American
- Consume Twitter regularly - from 4x/week to daily.
- Tweet often (2-3x/week)

These users can be found automatically via the Twitter API, but I anticipate volunteer response rate would be very low to cold-DM'ing them. Better to leverage the social networks of the people running the trial, family and friends, other students, and so on to attempt to recruit through retweets.

Proposed Treatment
In this experiment, I hope to leverage a feature that Twitter recently rolled out that allows users to bypass their default ranking algorithm and display their timeline in chronological order, rather than in the order of what Twitter considers the "best" tweets first. Study participants will be randomly assigned to treatment and control groups, and the treatment group will be instructed to change their timeline setting for a designated period of time (ie, two weeks). Because simply changing a setting may have some effect on consumption or creation patterns, control group will be instructed to change their Account time zone--this should have negligible effect on their Twitter browsing experience.

Primary Outcome: Survey Response
Prior to changing the settings, participants will be asked to complete a brief survey that describes their emotional state generally, and as it relates to Twitter. The questions should be adapted from a standardized anger instrument (ie, STAXI), perhaps modified to include some questions about social media usage patterns, and attempt to calibrate their level of anger both generally and specifically towards the topics they use Twitter to focus on (ie, politics, hobbies, friends). Participants would be asked to complete the survey again at the end of the trial, to determine if there is a difference.

Secondary Outcome: NLP Sentiment Analysis and Twitter Activity
Active Twitter users provide a window to their internal state--often their anger--through their interaction with the site. Users' historical tweets are publicly available, so it would be a simple matter to harvest their activity for the weeks preceding the trial, and compare it to their activity post-treatment. While the ideal scenario would be to train a custom sentiment analysis engine trained on Twitter to recognize anger, that might be outside the scope of this class. There are several sentiment-analysis algorithms freely available that could be applied to measure the positive or negative sentiment of the participants tweets. In addition, the analysis could be run

on the tweets that users have chosen to like and retweet, to see if changing the timeline has altered the signal that they choose to amplify to their followers.

Tertiary Outcome: Combination of the Two

People often mischaracterize their own behavior, may fail to recognize any changes, or may catch on to the purpose of the trial and allow that to bias their response. Normalizing the primary and secondary outcome responses to the same scale and then measuring the difference might provide the most comprehensive metric to evaluate a change.

For all outcomes, statistical analysis can consist primarily of comparing the distributions of participants' scores from before and after the trial, and measuring the p-value of the difference in means to determine level of significance.

Potential Risks

There are many opportunities for bias, noise and measurement error to be a danger in this endeavor. The sample population might be biased if we cannot recruit from a diverse enough source (eg, if we only recruit from friends of data scientists, we might get a very different response than if we sampled from the population as a whole). There may be inherent bias in the population of Twitter users themselves - that they are not accurately reflective of the United States population as a whole, and "Twitter beefs" do not extrapolate to larger societal tensions. Users might realize what the trial is about and consciously or subconsciously alter their behavior in response. Twitter is also a very noisy medium, and it can be hard to measure signal in the chaos. If there is a large-scale emotional event during the trial (eg, the Kavanaugh confirmation for Supreme Court)--and those events are fairly common--it could disrupt either of the outcome variables. There is also the potential for mismeasurement--"anger" is a relatively fuzzy concept to measure, and we would need to be careful that any change in emotional tone over the course of the trial is properly characterized. Finally, there is the risk that there is an effect but it is too small to be measured at this scale, and requires thousands of exposures over multiple years to accumulate enough to be noticeable.

Conclusion

That last risk seems pretty likely. In 2014, Facebook conducted a (now infamous) trial of nearly 700,000 users to determine whether deliberately rearranging their newsfeed to promote more positive or negative content to the top would affect their emotional state (as determined by their status updates). The effect they found was small, but their sample size was sufficiently massive that the change was statistically significant. This is similar to that (albeit at smaller scale and with user permission), but the key difference is the focus on anger, rather than a more vague notion of "positive" or "negative." The conventional wisdom these days is that the financial metrics of engagement incentivize social networks to write algorithms that surface content which infuriates users. Over the scale of years and thousands of exposures, that fury and resentment have built up to drive Americans further apart from each other. One key mechanic of this hypothesis is the ranking algorithm, and Twitter's recent settings change has provided a rare opportunity to test that algorithm outside of Twitter's offices.