INSTRUCTIONS FOR CODERS

1. Read the coding rules below.
2. Code the new dataset, based on the coding rules and your judgment.
1. E-mail the dataset back to me.

GENERAL INFORMATION

The datasets are in CSV (comma-separated values) format. They will open in Microsoft Excel or LibreOffice and have several columns:

1. class (**this column will be blank, and is where you code the tweet**)
2. Text: the full text of the tweet
3. user_bio_summary: the Twitter user's self-provided biography.
4. tweet_url: the full URLs of any links embedded in the tweet
5. username: the Twitter user's handle
6. real_name: the Twitter user's self-provided “real name”
7. user_location: self-provided user location
8. user_mention: “real names” of other Twitter users mentioned in a tweet
9. user_mention_username: Twitter handles of other users mentioned in a tweet (i.e., the handles corresponding to the user_mention list)
10. is_retweet: if non-zero, the tweet was a retweet; however, zero does not necessarily indicate that the tweet was not a retweet.
11. id (do not change this field)
12. id_text (do not change this field)

Coding is simply a matter of typing one of the three following values into the class column:

- “s” (without quotes) for Support
- “o” (without quotes) for Oppose
- “na” (without quotes) for tweets that cannot be coded either as support or oppose

Tip: double-clicking the dividing line between two column labels will automatically expand that column to accommodate the width of the widest cell in the column. However, doing this with some of the columns may take a long time.

Be sure to save your work often!

OVERVIEW

I have provided each coder with a dataset of 2400 net neutrality-related tweets. Of these, 1000 were randomly sampled from the complete dataset, and 1000 were selectively sampled from subsets of the complete dataset that, based on our research thus far, were likely to contain larger proportions of opposing tweets. The selective sampling was performed to boost the number of potentially opposing tweets in the sets we manually code, since the overall distribution in the complete dataset appears to be heavily lopsided in favor of net neutrality.[1]

Additionally, each dataset contains 200 tweets drawn randomly from each other coder's dataset. These

1 Since we are not using these samples to draw statistical inferences, they do not need to be representative. It is more important that they contain an adequate number of training tweets for poorly represented classes.

overlapping tweets will be used to assess inter-coder reliability.

CODING RULES: NET NEUTRALITY TWEETS

Coding tweets is difficult.

Use all the information available to you – I have tried to order the columns to provide the most useful information first. User-provided biographies can help illuminate the context of a tweet, as can the URLs linked to in a tweet.

Obviously, the tweet text itself is most important:

- RT @youngamer4con: The only way to ensure true #NetNeutrality is to minimize high barriers to entry, deregulate, and maximize competition i…
- RT @dandrabik: Really proud to see Obama fight for net neutrality. The internet should be regulated like a utility: http://t.co/KplSE2tWhA
- To those drinking the #netneutrality KoolAid let the #Gruber video be a lesson to you. The govt has to lie in order to sell it to you.
- My problem with #NetNeutrality is I don't want large monopolies or the government in control of the Internet... https://t.co/ft3bOOVKFi

The first two tweets appear to support net neutrality, while the second to tweets appear to oppose it.

However, not all tweets are as easy to interpret:

- @CutGovtSpending You're still wrong about net neutrality though :)
- RT @Jantxnc: Net Neutrality: A 21st Century Trojan Horse http://t.co/DoebTXZtf5 via @townhallcom

With both of these tweets, the text itself is somewhat ambiguous. We might guess that a tweet directed at an account called “CutGovtSpending” and saying that the account is “wrong” might support net neutrality, and we might guess that calling Net Neutrality a “Trojan Horse” might oppose it, but it's difficult to say for certain.

In these cases, tweet and user metadata can help us code the tweet. **You can and should use the additional information to help you code the tweet.**[2]

For example, the following is a real user biography:

- When Tyranny Becomes Law, Rebellion Becomes Duty. #UniteRight #SueLiberals. Oppose:#WarOnChristians #WarOnWhites Support:#TCOT #TeaParty #2A #Militia #OCRA #GOP

This biography can help you interpret a tweet that by itself might appear ambiguous.

Likewise, here is an example of a real URL list:

2 This additional information will also be included in the features provided to the machine learning algorithms.

- [u'http://reason.com/blog/2014/11/13/obamas-scheme-to-regulate-us-into-broadb']

It is a link to Reason.com, which by itself tells us information about the Twitter user; furthermore, the remainder of the URL contains important information about the nature of the tweet.

Keep in mind that we are coding for support or opposition to the public idea of net neutrality. Thus, you may encounter tweets opposing the FCC's hybrid plan, but ultimately supporting the concept of net neutrality. On the other hand, you may encounter tweets opposing the FCC's hybrid plan on the grounds that it is government interference in the market, which suggests the user opposed the concept of net neutrality.

Here are some general guidelines to give you a sense of how to make coding decisions.

Many of these guidelines distinguish between "bare" retweets and retweets with added commentary. In the latter case, **you should always used the added commentary** to interpret the meaning of the tweet. This goes even for tweets not captured by the guidelines below.

CODE AS SUPPORT ("s"):

- pro-net neutrality tweets
- **Retweets of pro-net neutrality tweets**
- **Retweets of news stories about net neutrality**
  - **unless they have added commentary by the user (in which case use this commentary as the determining factor)**
  - **or the news story itself is clearly opinionated (in which case use the opinion as the determining factor)**
- **Tweets critical of the FCC's hybrid plan (or Tom Wheeler) *on the grounds that it doesn't go far enough***
- Retweets of Oatmeal comic about net neutrality
- Tweets expressing support for Obama/White House plan in general
- Retweets of White House tweets about net neutrality
  - unless they have added commentary...
- Tweets criticizing large ISPs like Comcast/Verizon
- Tweets attacking or denigrating conservatives/GOP/Republicans/Ted Cruz in connection to net neutrality
- tweets that ask to reclassify Internet service providers (ISPs) as "common carriers," "utilities," a "telecommunications service," or "public utilities"
- tweets that ask the FCC to exercise its Title II authority

CODE AS OPPOSE ("o"):

- anti-net neutrality tweets
- **Tweets using the #tcot (top conservatives on Twitter) hashtag**
  - **unless using the hashtag in an ironic/sarcastic manner)**
- **Tweets critical of the FCC *on the grounds that it goes too far***
- **Retweets of Ted Cruz's "Obamacare for the Internet" tweet**
  - **unless they have added commentary by the user (in which case use the commentary)**
- Tweets opposing government interference, regulation, control, censorship etc., of the Internet

- Tweets expressing concern about the effect of net neutrality on investment or competition
- Tweets expressing concern about the effect of net neutrality on freedom/liberty
- Retweets of Mark Cuban's tweets about net neutrality
  - unless they have added commentary...
- Tweets expressing opposition to Obama/White House plan in general
- Tweets tying net neutrality to Benghazi/Gruber/Obamacare
- Tweets attacking or denigrating liberals/Democrats/Obama in connection to net neutrality

CODE AS UNCODEABLE (“na”):

Do your best to avoid coding tweets as uncodeable. Uncodeable tweets will simply be discarded, since the machine learning techniques we will be using perform best with two classes. However, the computer will still encounter ambiguous tweets, and will assign a class to them. The more information we can provide to the computer about how to deal with these ambiguous tweets, the better.

That said, there will be some tweets that simply cannot be assigned to either the support or oppose class:

- tweets where there simply is not enough information to make a reasonable guess about support or opposition
- tweets that are in foreign languages (**even if you can read them**)
- tweets that are spam or advertisements for unrelated products/services

Overall these are guidelines, not rules. You'll have to use your judgment and general knowledge in many cases.