Quantitative Manual Content Analysis

Text as Data

Amber Boydstun & Cory Struthers

Spring 2025

Contents

Introduction . . . 1
Lab objectives . . . 1
1. Stop and think about your measurement goals! . . . 1
2. Look at your data . . . 2
3. Build and test a codebook . . . 4
4. Code your data independent from another coder . . . 4
5. Check for inter-coder reliability . . . 4
6. Wash, rinse, and repeat . . . 4
Homework . . . 5

Introduction

Manual content analysis is the gold standard for understanding—and validating—all computational approaches to text analysis.

Lab objectives

- Stop and think about your measurement goals!
- Walk through the joys and challenges of manual content analysis
- Calculate inter-coder reliability
- Understand the value of human reading of text

1. Stop and think about your measurement goals!

Before you do anything, with any text, Step 1 is to think hard about your measurement goals, namely your research question and exactly what you want to measure. Your measurement goals should guide all other decisions. This is true for manual content analysis but also for text-as-data approaches of all kinds.

What is your research question? Do you want to explore some text, without strong expectations of what it holds? Do you want to test a pre-formulated hypothesis about specific patterns you expect to find in the text? Do you want to use the text to measure a key variable that you will then use alongside other metrics for a bigger project? Whatever your motivations are for picking up a body of text and analyzing it, you want to have those motivations solid—preferably in writing—before you begin. The land of text as data is a magical land, but one filled with many rabbit holes where one can easily get lost.

What is your latent variable(s) of interest? Unless your goal is to explore a body of text without any expectations of what you will find, you should think conceptually about what variable(s) you want to measure. Maybe you're interested in comparing positive and negative language between different news sources, or examining whether the level of emotion expressed in presidential speeches has changed over time,

or assessing how similar different high school history texts are. Whatever your research goal, it's important to think long and hard about the core thing (i.e., variable) you want to measure and the pros and cons of different approaches to measuring it.

What population of text should you use? For example, if you're interested in the discussion of pets by young adults on social media in the last year, you probably don't want to look at Facebook, or even Twitter. You probably want to look at Instagram, Snapchat, and/or TikTok, and you might want to subset your population to only those posts authored by people under a certain age. However, if you're interested in comparing how different people of different ages talk about their pets on social media, you'd likely want to include all users across all these platforms (and more).

What unit of analysis should you use? An often overlooked but *crucial* question is what unit of analysis you should use. In many cases, the unit of analysis is the **document**. For instance, some researchers may want to measure how many times each newspaper article or social media post references a term like "liberal", "conservative", or (better yet) "giraffe". However, in some cases the unit of analysis might be smaller or larger than the document: the analyst may want to break apart a newspaper article into paragraphs, or combine all the tweets written by a single author in one day. In our example about discussion of pets, for example, should you use treat each social media post as a unit (i.e., treat it as a separate document)? Or would it be better to break up the posts into sentences, with the sentence as the unit of analysis? Or should you expand it out, pooling all the posts by month, and treating the month as your unit of analysis, with all posts on that month treated as one big document? There are no wrong answers here. Which unit of analysis you should use depends on—you guessed it—your research question/goal and your latent variable(s) of interest.

2. Look at your data

Once you've made these important decisions, the next step for manual content analysis is to LOOK at your data, keeping your research goal and latent variable(s) of interest in mind. Let's use a sample of Twitter posts about immigration as an example. For this exercise, let's use the tweet as our unit of analysis.

In order to look at our data in R, we will need the following packages.

We also need to set our working directory to the local or remote folder that hosts the data required for this course to use relative paths when loading data. Time constraints prevent us from diving deeply into differences between absolute and relative paths, but this source provides helpful background. Relative paths are considered best practice for collaborative projects, so that is what we'll set up here using `setwd()`. We will use the same line of code every time we begin a new module. Note that on Windows, we use / (forward slash) instead of (back slash). Then we can load the data and take a peek at it.

```
# Load packages
require(tidyverse)
require(dplyr)

# Set working directory
setwd("~/Library/CloudStorage/Box-Box/A_teaching/POL290G_TextAsData/modules/data/")
#getwd() # view working directory

# Load data
tweet_data = read.csv("sample_immigration_tweets_2013-2017.csv")
class(tweet_data)
```

```
## [1] "data.frame"
```

```
# Look at your data
head(tweet_data)
```

```
##      tweet_id year month day hour minute second is_retweet retweets
## 1 9.20638e+17 2017    10  18   13     11      5          0        0
```

```
## 2 9.01247e+17 2017     8  26    0     56     51          0        0
## 3 7.97872e+17 2016    11  13   18     43      8          1        0
## 4 8.25354e+17 2017     1  28   14     45     11          1        0
## 5 8.24631e+17 2017     1  26   14     50     42          1        0
## 6 8.28766e+17 2017     2   7    0     43     23          1        0
##                                                  user_url follower_count
## 1            http://steamcommunity.com/id/gayskeleton1992/            385
## 2                              http://unitedresistance.com/           1642
## 3                                                     <NA>            206
## 4                                                     <NA>             58
## 5                                                     <NA>             30
## 6 https://www.youtube.com/channel/UCJNjv114NMZ-ja76ErSB-rw           1723
##   verified
## 1    FALSE
## 2    FALSE
## 3    FALSE
## 4    FALSE
## 5    FALSE
## 6    FALSE
##
## 1                                                                 Obama once called me and he said
## 2                    Trump Pardons Sheriff Joe Arpaio - Mr. Arpaio, who built a national reputation
## 3                                                               RT @AFP: #BREAKING Trump vows to
## 4 RT @PersianCeltic: It''s official America a country founded by immigrants, after stealing land fro
## 5    RT @RobSchneider: My mother is an immigrant from the Philippines. My grandfather was a Jewish i
## 6          RT @AnnCoulter: If GOP is worth anything, they'd impeach wacko Judge Robart, who claims t
##          Anti     Neutral         Pro Capacity.and.Resources
## 1 0.001030748 0.998835927 0.000133325              5.62e-06
## 2 0.999595635 0.000266326 0.000138039              2.45e-05
## 3 0.999681621 0.000250103 0.000068300              2.36e-05
## 4 0.000044100 0.000086800 0.999869059              4.94e-05
## 5 0.000029200 0.000086300 0.999884438              1.02e-05
## 6 0.233192650 0.765425231 0.001382118              1.76e-05
##   Crime.and.Punishment Cultural.Identity    Economic
## 1          0.000014600        0.00002450 0.000009490
## 2          0.998977054        0.00005090 0.000029000
## 3          0.006385019        0.00007730 0.000048600
## 4          0.000073900        0.04255607 0.000124714
## 5          0.000009720        0.99970370 0.000011100
## 6          0.000256318        0.00003030 0.000043400
##   External.Regulation.and.Reputation Fairness.and.Equality Health.and.Safety
## 1                        0.000009440           0.000115101          1.14e-05
## 2                        0.000027900           0.000093800          3.57e-05
## 3                        0.000044700           0.000188458          4.45e-05
## 4                        0.000111604           0.955533214          8.47e-05
## 5                        0.000012100           0.000053400          7.07e-06
## 6                        0.000027800           0.000136101          3.19e-05
##   Legality..Constitutionality..Jurisdiction    Morality    Other
## 1                              0.000041600 0.000013700 2.65e-05
## 2                              0.000195348 0.000321173 1.56e-05
## 3                              0.000134543 0.000052700 3.38e-05
## 4                              0.000048100 0.000230013 6.97e-05
## 5                              0.000015000 0.000036400 1.27e-05
## 6                              0.002767632 0.000073200 4.07e-05
```

```
##   Policy.Prescription.and.Evaluation   Political Public.Sentiment
## 1                        0.000021500 0.999683616         5.74e-06
## 2                        0.000031100 0.000099500         3.77e-05
## 3                        0.000216373 0.992617148         3.64e-05
## 4                        0.000263481 0.000418053         4.29e-05
## 5                        0.000015400 0.000051700         7.52e-06
## 6                        0.000040500 0.996453785         2.83e-05
##   Quality.of.Life Security.and.Defense
## 1     0.000009630             7.60e-06
## 2     0.000029300             3.14e-05
## 3     0.000043400             5.34e-05
## 4     0.000310655             8.35e-05
## 5     0.000038400             1.55e-05
## 6     0.000027300             2.52e-05
```

Now we can think about a key variable we'd like to measure, and scan through the tweets to get a sense for how easy it would be to categorize each tweet according to your variable. For a real project, we'd want to take a random sample of the text and read it at this stage.

3. Build and test a codebook

With your variable in mind, you'll then want to write down the rules that will allow you (and future researchers) to categorize each text according to that variable. For example, if we want to track social media posts about pets, we'll want to establish a rule telling us whether or not someone's post about a *friend's* pet should count.

4. Code your data independent from another coder

When you've established your codebook, you'll want to have two people (coders) independently use the codebook to code the same random sample of data in order to see whether the codebook alone is thorough enough to produce consistent coding decisions.

```
# Subset to only 20 observations and only the text variable
tweet_df = subset(tweet_data, select = c(text)) %>%
  slice(100:120)
head(tweet_df)

# Add two columns for your own coding: one for valence, the other for a variable you find interesting
tweet_data_coded = cbind(tweet_df, valence="", other_variable="")
head(tweet_data_coded)

# Output your data for manual coding (writing into your current working directory folder)
write.csv(tweet_data_coded, "~/Library/CloudStorage/Box-Box/A_teaching/POL290G_TextAsData/modules/data/
```

5. Check for inter-coder reliability

Use Deen Freelon's extraordinary (and free!) online system to calculate inter-coder reliability: http://dfreelon.org/utils/recalfront/

In general, a Cohen's Kappa and Krippendorff's Alpha scores of over 0.7 allow us to be confident in the reliability of the coding.

6. Wash, rinse, and repeat

You'll want to iterate through Steps 1-5 until you have a solid concept of your variable of interest, a strong codebook that can handle most new observations you code, and high levels of inter-coder reliability. Make

sure to annotate your codebook as you go, making notes of any specific coding decisions you make. But careful! If you make a decision that is inconsistent with how you have coded things in the past, you'll need to go back and find those observations and re-code them according to your new rule. Fun, right?

Homework

Discussion Questions:

1. What signs should you look for to know that your codebook is ready and doesn't need more editing?
2. Beyond the best practices already discussed, what other best practices would make sense to employ in using quantitative manual content analysis?

Project Questions (show your work): Apply this approach to your own research/interests!

1. Work with your partner or team to develop a single variable of interest and a brief codebook, using a corpus of your choice. What was the process like? What were the easy parts, and what were the challenges?
2. Subset your corpus to 20 observations, and have each partner code those 20 items independently. What was the process like? How confident did you feel in your coding, and what did you observe from spending time with the text?
3. Calculate your inter-coder reliability (follow Professor Freelon's instructions about how to format the .csv file!). What factors help explain how high or low it is? Which particular observations did you code differently, and how might you update the coding process to improve the inter-coder reliability?
4. Now repeat steps 1-3 again, but using a different variable of interest OR a different set of coding rules for your same variable of interest.
5. Finally, discuss: what worked and what didn't, and why? Explain the similarities and differences between your two coding approaches and what you learned.