Instructions for creating “consensus” categories

Version 6, 23 February 2022

Background

We are analyzing a dataset of texts in order to identify categories of things being discussed. Examples of texts that can be analyzed in this way might include sets of open-ended text survey responses, social media posts, conversational turns in interview transcripts, or public comments to a federal agency.

The analysis we’ve done so far involves two steps. First, there was an automatic computational analysis of the whole dataset to produce “topics” (sometimes also called “themes”) that represent categories or concepts found in the collection of text. This first step is fast and automatic, but it doesn’t give names or labels to the topics that were found, and it also doesn’t provide useful, readable descriptions of what the categories are about.

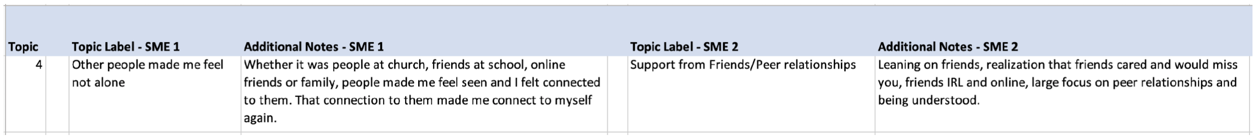
In the second step, two subject matter experts looked independently at the automatic analysis from the first step. They both followed a set of step-by-step instructions in order to assign a name or label to each category, along with writing down free-form descriptions and additional notes about each one to describe what the category was about. If you have prior experience with coding and codeset development, you can think of what they did as a process of codeset discovery, in order to independently produce a set of meaningful codes.

Objective

The next step in the process is to look at what the two subject matter experts (SMEs) produced independently, and combine their results into a single “consensus” analysis of the categories that were discovered. Sometimes we ask the original SMEs to work together to do this, sometimes we ask someone new to do this part, and sometimes we ask two new people to work together. By the time you’re reading this, we should have already let you know what the plan is.

More specifically, the spreadsheet you’re getting contains the two independent sets of category labels along with corresponding descriptions/notes. Based on these, for every category, the goal is to come up with a single name or label that captures the topic or concept, and to provide a free-form description or notes including any compromises or considerations you made during the process. Optionally, as discussed below, you may also choose to group the topics into higher-level categories.

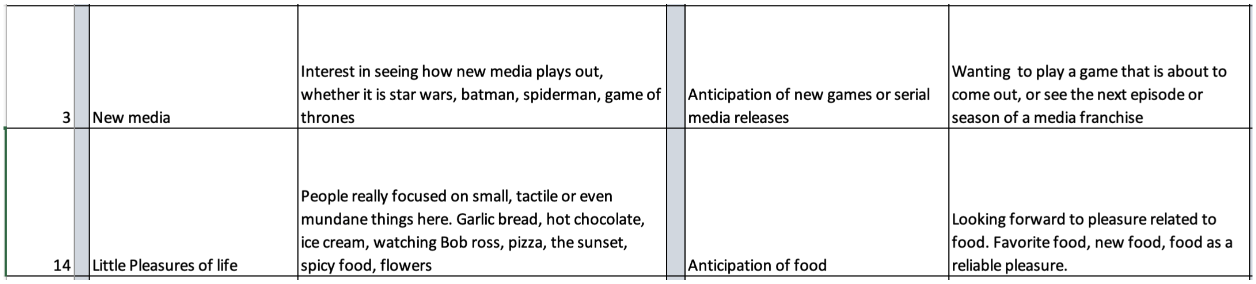
As an example, in a previous study, we analyzed writing by people related to their mental health, talking about how they managed to get through difficult times. Here is the analysis for topic category #4 in that study by two independent mental health subject matter experts.



Topic	Topic Label - SME 1	Additional Notes - SME 1	Topic Label - SME 2	Additional Notes - SME 2
4	Other people made me feel not alone	Whether it was people at church, friends at school, online friends or family, people made me feel seen and I felt connected to them. That connection to them made me connect to myself again.	Support from Friends/Peer relationships	Leaning on friends, realization that friends cared and would miss you, friends IRL and online, large focus on peer relationships and being understood.

A good consensus label for this topic might be “Sense of Connection with Others”, and a consensus description/notes might be “Sense of connection -- feeling seen, heard, understood, cared about by other people”.

Here are two more examples.



3	New media	Interest in seeing how new media plays out, whether it is star wars, batman, spiderman, game of thrones	Anticipation of new games or serial media releases	Wanting to play a game that is about to come out, or see the next episode or season of a media franchise
14	Little Pleasures of life	People really focused on small, tactile or even mundane things here. Garlic bread, hot chocolate, ice cream, watching Bob ross, pizza, the sunset, spicy food, flowers	Anticipation of food	Looking forward to pleasure related to food. Favorite food, new food, food as a reliable pleasure.

A consensus label for the first one might be “Looking forward to next episodes of games and media”, and for the second one it might be “Food and sensory pleasures”.

Optionally, there will be the opportunity to create a higher-level category that contains multiple topics. For instances, in the mental health study the people doing what you’re doing decided that topics #3 and #14 in the above example, together with some other topics, group naturally into a higher-level category they called “Activities or events that provide enjoyment, distraction, anticipation, an alternative focus”.

Updates to the spreadsheet format. Note that in the materials you’ll be working with, the format of this spreadsheet may have changed from what you’re seeing in the examples above. In particular, we’ve been updating the spreadsheet described above to look a little more like a traditional codebook by having one column for a category’s official description, and a separate column for any additional notes or scratchwork. We have also added a “Coherence” column where the independent SMEs have provided a 0-to-3 rating for each automatically-discovered category, which we’ll describe further below.

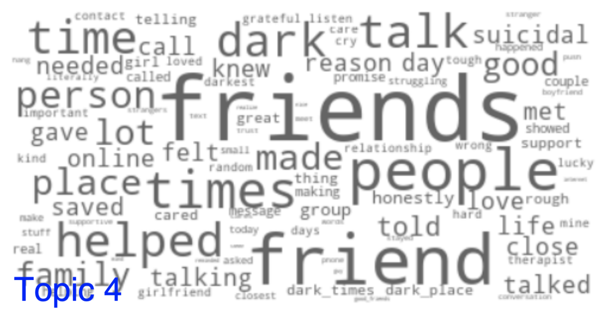
The materials you’ll be working with

Labeling spreadsheet from the two independent subject-matter experts (SMEs). Along with these instructions, you should have received a spreadsheet similar in format to the example above. This is the main file you’ll be working with. You’ll see that to the right of the labels and notes the previous SMEs provided, there are two new columns for you to fill in – *Consensus*

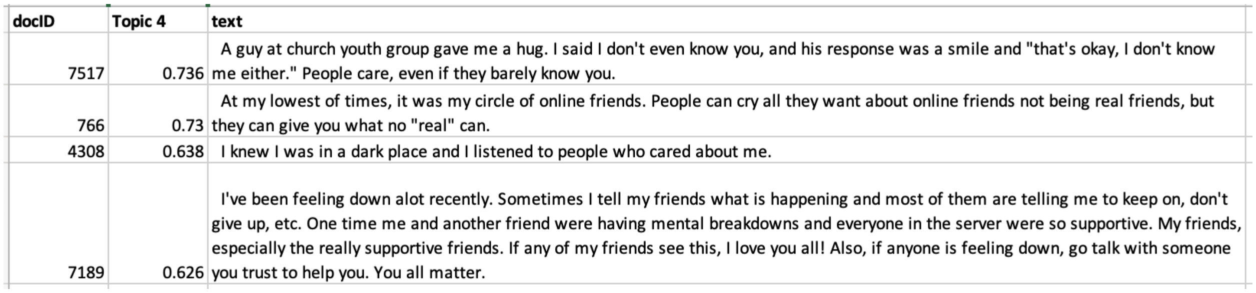
Topic Label and *Consensus Additional Notes* – plus there are optional *Higher Level Category* and *Higher Level Notes* columns.

Additional materials you might find useful. The labeling spreadsheet should be all you need. However, it's possible that in some cases you might need to look at other materials – in particular, the materials that the subject matter experts looked at when they formed their opinions about the categories. They looked at two sources of information from the automatic computational analysis:

"Clouds" of words that help indicate what a topic is about. One thing they looked at was a PDF file containing "clouds" of words for each topic. For example, here is the set of words they saw for topic category #4 from the example above that led them to think it might be about friends and connection with others (where, the bigger the word is, the more strongly it helps indicate what the topic is about):



Topic scores for each "document", i.e. each piece of text. In addition, they looked at a big spreadsheet containing all the texts. In that spreadsheet, each discovered topical category gets a column, and the numerical "score" in the column represents how strongly the category is present in that piece of text. For example, if you use Excel to re-order the rows in the spreadsheet using the scores for that topic category #4, highest scores at the top, you see this in the top rows:[1]



docID	Topic 4	text
7517	0.736	A guy at church youth group gave me a hug. I said I don't even know you, and his response was a smile and "that's okay, I don't know me either." People care, even if they barely know you.
766	0.73	At my lowest of times, it was my circle of online friends. People can cry all they want about online friends not being real friends, but they can give you what no "real" can.
4308	0.638	I knew I was in a dark place and I listened to people who cared about me.
7189	0.626	I've been feeling down alot recently. Sometimes I tell my friends what is happening and most of them are telling me to keep on, don't give up, etc. One time me and another friend were having mental breakdowns and everyone in the server were so supportive. My friends, especially the really supportive friends. If any of my friends see this, I love you all! Also, if anyone is feeling down, go talk with someone you trust to help you. You all matter.

[1] For a 2-minute video on how to sort (i.e. re-order) the rows in your spreadsheet by the values in one of the columns, see https://youtu.be/9KjkVDH3_ig. For a one-minute video on how to hide and unhide columns in Excel, so that you can look at a smaller number of columns at a time, see https://youtu.be/trk1MIOynm8. For a one-minute video on how to "wrap" long text in Excel, so that instead of the text just going outside the edges of the cell, instead the cell will expand to fit all the text into it, see https://youtu.be/CiWjGKXvrbI. It can be helpful to make a column wider first, and then "wrap" for easier readability.

Note that the independent SMEs have already done their work creating labels and descriptions for the topic, and your job here is not to do the same work over again. You should only look at the clouds and the document spreadsheet *if you need to* in order to help you understand what the previous SMEs were getting at with particular topics, e.g. if their labels or notes are very unclear or hard to understand or if they seem to conflict significantly with each other. Otherwise please work from the combined spreadsheet, not the original materials. We recommend deciding in advance on a maximum amount of time you'll spend per topic, and setting a timer when you start working on each topic. In past experience we've found that 2 to 5 minutes per topic on average is sufficient to arrive at good results.

What you should do

First, please keep track of the time you spend working on this. We don't expect that it should take you more than a few hours, but knowing how long you actually took will be useful information for other studies.

Please follow these steps:

1. Open the spreadsheet with the combined results from the two independent SMEs, which will be called *combined.xlsx* or *sme_combined.xlsx* or something similar.

2. Go through the topics one by one. For each one:

    a. Read the topic labels, descriptions, coherence scores, and associated notes for this topic.
    b. See if you can construct a single label and a single description/notes to capture the consensus of what the two SMEs were seeing in this category, informed by that information.
    c. If it's hard to understand what the SMEs were getting at, or if they seem like they're significantly in conflict with each other:
        i. Open the PDF "clouds" file and the document-topics Excel file.
        ii. Re-order the rows so that the scores for this topic go from largest to smallest (see Footnote 1 for guidance on how to re-order rows and do other useful things in Excel).
        iii. Spend *no more than a minute or two* looking at the texts in the top rows to form your own understanding of what this topic might be about, informed by what the previous SMEs thought. Setting a timer for 1 to 2 minutes is a good idea; otherwise it's easy to get pulled into too much reading, because often the texts are very interesting!
        iv. Finally, see if you are *now* able to construct a single label and a single description that captures what the two SMEs were seeing in this category.
        v. Update the *Consensus Additional Notes* column with notes about the difficulty you had and what you did about it.

d. If you are still unable to arrive at a single consensus label and description that seem reasonable, use the label UNSURE and write a brief description of the problem you're having. For example, a description might be "No consensus: the two labels and descriptions are clear, but they're so different I can't find any way to combine them". Or another possible description might be "I just can't make sense of what the 1st SME's label and notes are trying to say".

e. Optionally, if you are seeing that some topics may group together naturally into higher-level categories, you can make notes about that in the column labeled *Optional: Higher Level Group Notes* as you go, and then in step 3 (below) you can choose, if you like, to create higher-level group labels.

3. Optionally, after you've gone through all the topics, fill in the two columns *Optional: Higher Level Group Label* and *Optional: Higher Level Group Notes*. (You may already have notes in the latter column; see step 2e above.) The illustration above using topics #3 and #14 provided an example where two topics with consensus labels "Looking forward to next episodes of games and media" and "Food and sensory pleasures" could both be assigned a higher-level category you might name "Activities or events that provide enjoyment, distraction, anticipation, an alternative focus". Note that it is possible none of the topics can fit into a higher-level grouping, or maybe only some of them do. If some topics are grouped into higher level categories and others aren't, that's ok.

4. Send us back the spreadsheet where you've filled in the new consensus columns. Please also copy/paste the following questions into your cover email and provide answers:
    a. How long did you spend on the entire process, including the instructions (but leaving out any breaks you took)?
    b. Did you do it all in one sitting or did you take breaks?
    c. On a scale from 1 (terrible) to 5 (great), how was the quality of the materials you were given to work with?
    d. If you've engaged in more traditional codeset development before, please comment on how this compares with that previous experience.
    e. Please identify any aspects of the process that you particularly liked.
    f. Please identify any aspects of the process that you think could be improved, and we welcome any specific suggestions for improving it.

Thank you for your help on this project!