‘When I’m retired...’

Using topic modeling to analyze open-ended survey questions in the A Broader Mind Longitudinal Survey

Arjen de Wit

December 2020

Besides the standard survey instruments that are based on closed-ended items, open-ended questions can provide rich data. In this document I will analyze the answers to an open-ended survey question in which respondents are asked to imagine themselves in the future. Data are used from the A Broader Mind Longitudinal Survey, an annual survey that tracks the civic, social, professional and academic development of students at Vrije Universiteit Amsterdam. See a pilot report [here](#). For the analysis below I use data from the 2018 pre-test sample, which consists of European university students (n=200) sampled through the online platform Prolific.

These analyses are meant to explore the potential of including analyzing open-ended questions in our survey. The rationale behind the survey is that students develop themselves not only in an academic and cognitive sense, but also as persons in a complex society. To what extent do students, during their time at the university, adopt ‘a broader mind’ on themselves, others and societal issues?

The survey was designed together with Rosan van Niekerk and René Bekkers. At the end of the survey, we asked respondents to imagine the things they would like to accomplish in their future lives. This provides another measure of one’s views on the future, besides the closed survey questions that consume the largest part of the questionnaire. We randomized the wording of the question to examine whether a different time span (in 20 years vs. when you are retired) and a different qualification (proud vs. satisfied) would lead to different answers.

The question asked in the survey was: “(A.1) If you think about yourself and your career (A.2), when would you be (B)? Please elaborate on your answer.”

- 25% of respondents received A1: ‘Imagine you are retired’ and B: ‘proud’;
- 25% received A2: ‘in 20 years’ and B: ‘proud’;
- 25% received A1: ‘Imagine you are retired’ and B: ‘satisfied’;
- 25% received A2: ‘in 20 years’ and B: ‘Satisfied’.

Below I will show how Latent Dirichlet allocation (LDA) distinguished three topics in the open answers, reflecting the importance of (1) a steady job, (2) sufficient time for family, and (3) a contribution to society. The formulation of the open question did affect the answers. ANOVA tests revealed that topic 1 was significantly less often mentioned in the retired condition ($F(1)=6.628$, $p=0.011$), while topic 3 was more often mentioned in this condition ($F(1)=12.480$, $p=0.001$). Using the word ‘proud’ vs. ‘satisfied’ did not yield significantly different answers.

Disclaimer: I am not an expert on the issues covered here. As a relative beginner in R and text mining I built on many great open access resources, most importantly Silge & Robinson’s Text Mining with R. This document explores the possibilities of using topic modeling in a specific survey. I hope this contributes to improving survey design and survey research, even for researchers in the social sciences who are not necessarily methodologists.

1. Data preparation

The data was collected in Qualtrics. After downloading, I recoded most variables and anonymized the data. This is the dataset I start with.

```
abmls <- readRDS(params$filename)
```

Answers in the four experimental conditions are each in a separate column of the data. First, I merge them into one column. I make a new data frame with only the text:

```
library(dplyr)
abmls$open <- coalesce(abmls$Q49, abmls$Q50, abmls$Q52, abmls$Q51)
open <- data_frame(line = 1:201, text = abmls$open)
```

```
open
## # A tibble: 201 x 2
##   line text
##   <int> <chr>
## 1      1 I would be satisfied if I feel that I have made a positive contri~
## 2      2 if i could make a nice living
## 3      3 I will be satisfied when I have a steady income, am surrounded by~
## 4      4 If I have lived my life more than I have worked. My main goal is ~
## 5      5 Thinking ahead 20 years from now, I will be proud of myself when ~
## 6      6 I would be proud, if I managed to reach what was really hard to a~
## 7      7 A job with high responsibility in my field, earning a salary whic~
## 8      8 When I lived happy life
## 9      9 When I had my university degree in psychology and had the chance ~
## 10     10 I'd like to be self-sufficient and in some kind of teaching role,~
## # ... with 191 more rows
```

We see that each row is an answer, often with multiple sentences. For text analysis we want to have a tidy dataset, in which each word is in a separate row:

```
library(tidytext)
tidy_open <- open %>%
  unnest_tokens(word, text)
```

```
tidy_open
## # A tibble: 6,306 x 2
##   line word
##   <int> <chr>
## 1      1 i
## 2      1 would
## 3      1 be
## 4      1 satisfied
## 5      1 if
## 6      1 i
## 7      1 feel
## 8      1 that
## 9      1 i
## 10     1 have
## # ... with 6,296 more rows
tidy_open %>%
```

```
count(word, sort = TRUE)
## # A tibble: 1,070 x 2
##   word      n
##   <chr> <int>
## 1 i      494
## 2 to     263
## 3 a      237
## 4 and    226
## 5 be     174
## 6 my     158
## 7 would  141
## 8 have   134
## 9 in     132
## 10 the   109
## # ... with 1,060 more rows
```

We can see that the most common words are 'I', 'to', 'a', 'and', etc. These are not so meaningful so we want to exclude them. Additionally I excluded five words that are specific to the question asked, because respondents often repeat a part of the question in their answers: 'proud', 'satisfied', '20', 'career' and 'life'.

```
stop_words2 <- stop_words
stop_words2$lexicon <- NULL
stopwextra <- c("proud", "satisfied", "20", "career", "life")
stopwextra <- data_frame(line = 1:5, text=stopwextra)
stopwextra <- stopwextra %>%
  unnest_tokens(word, text)
stopwextra$line <- NULL
stopw <- rbind(stopwextra, stop_words2)
tidy_open <- tidy_open %>%
  anti_join(stopw)
```

```
tidy_open %>%
  count(word, sort = TRUE)
## # A tibble: 760 x 2
##   word      n
##   <chr> <int>
## 1 job      45
## 2 family   37
## 3 people   29
## 4 money    26
## 5 time     26
## 6 happy    25
## 7 feel     21
## 8 society  21
## 9 income   18
## 10 live    17
## # ... with 750 more rows
```

We can see that the most common words are 'job' (occurs 45 times) and 'family' (37 times). To get an overview of the most common words, I make a bar chart of the words that occur more than 10 times in the data:

```
library(ggplot2)
```

```
tidy_open %>%
  count(word, sort = TRUE) %>%
  filter(n > 10) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```

I make a new dataframe with a column for the respondent (resp) and the number of times the word occurs.

```
tidy_openn <- tidy_open %>%
  count(word, sort = TRUE)
tidy_open$resp <- tidy_open$line
tidy_open <- full_join(tidy_open, tidy_openn, by = "word", copy=FALSE)
```

```
tidy_open
## # A tibble: 1,639 x 4
##   line word      resp    n
##   <int> <chr>    <int> <int>
## 1     1 1 feel          1    21
## 2     2 1 positive       1    13
## 3     3 1 contribution  1     3
```

```
## 4      1 society      1    21
## 5      1 lived       1     7
## 6      1 pursued     1     1
## 7      1 society     1    21
## 8      2 nice        2     7
## 9      2 living      2     4
## 10     3 steady      3    10
## # ... with 1,629 more rows
```

Then I make a Document-Term Matrix.

```
open_dtm <- tidy_open %>%
  cast_dtm(resp, word, n)
```

2. Identifying topics

To identify the topics in these data, I used Latent Dirichlet allocation (LDA). In LDA, you have to specify the number of clusters k that you want to have identified. I tried the model multiple times with different number of clusters. By plotting the ‘best fitting’ words for each cluster, we can interpret the clusters theoretically and judge whether these topics make sense.

```
library(topicmodels)
open_lda <- LDA(open_dtm, k = 3, control = list(seed = 1234))
open_topics <- tidy(open_lda, matrix = "beta")
open_topics
```

```
open_top_terms <- open_topics %>%
  group_by(topic) %>%
  top_n(20, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
open_top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```

As you can see, I chose a solution with three clusters. Each word has a *beta* score for each cluster, which is the probability of a word coming from that topic.

The first cluster refers to a steady job and a stable life, with words like ‘job’, ‘income’, ‘live’ and ‘stable’ as top terms. The word ‘happy’ suggests that respondents not only want a lot of money, but directly relate this to a happy life. The second cluster relates to having sufficient time for family, with words like ‘family’, ‘time’, ‘enjoy’ and ‘support’. ‘Money’ may refer to having sufficient resources to support one’s family. The third cluster is more than the others related to contributing to society. Terms with high *betas* here are ‘people’, ‘society’, ‘impact’ and ‘positive’.

3. Topic probabilities

Each open answer (‘document’) has a topic probability *gamma* for each topic, which is the estimated proportion of words in that answer which belongs to that cluster. I made a wide dataframe with these gammas for each topic, in which the document number corresponds with the id number of each respondent.

```
open_documents <- tidy(open_lda, matrix = "gamma")
library(reshape2)
open_documents_wide <- dcast(open_documents, document ~ topic, value.var="gamma")
open_documents_wide$topic1 <- open_documents_wide$`1`
open_documents_wide$topic2 <- open_documents_wide$`2`
open_documents_wide$topic3 <- open_documents_wide$`3`
open_documents_wide$`1` <- NULL
open_documents_wide$`2` <- NULL
open_documents_wide$`3` <- NULL
```

```
open_documents_wide$id <- open_documents_wide$document
open_documents_wide$document <- NULL
```

Let's examine these *gammas*.

```
library(ggplot2)
library(gridExtra)
```

```
p1<- ggplot(open_documents_wide, aes(open_documents_wide$topic1)) +
  geom_histogram(binwidth=.02) + xlab("Gamma topic 1")
p2<- ggplot(open_documents_wide, aes(open_documents_wide$topic2)) +
  geom_histogram(binwidth=.02) + xlab("Gamma topic 2")
p3<- ggplot(open_documents_wide, aes(open_documents_wide$topic3)) +
  geom_histogram(binwidth=.02) + xlab("Gamma topic 3")
grid.arrange(p1, p2, p3)
```

Although no individual score is exactly 0 or exactly 1, most values are at the extreme ends of the scale. The large majority of *gammas* are very close to 0. I think this is due to the small size of the 'documents': when using, say, 40 words, it is likely that an answer can be identified with only one topic.

4. Survey experiment

After having identified the topics, I merged the LDA results with the original dataset. This allows for correlating the topic probability *gamma* with the experimental conditions and other covariates in the data.

```
abmls$id <- rownames(abmls)
abmls <- full_join(abmls, open_documents_wide, by = "id", copy=FALSE)
```

I made a variable that identifies the experimental condition of each respondent.

```
abmls$opencond <- NA
abmls$opencond[!is.na(abmls$Q49)] <- 1 #20/PROUD
abmls$opencond[!is.na(abmls$Q50)] <- 2 #20/SATISFIED
abmls$opencond[!is.na(abmls$Q52)] <- 3 #RETIRED/PROUD
abmls$opencond[!is.na(abmls$Q51)] <- 4 #RETIRED/SATISFIED
table(abmls$opencond)
abmls$openretire <- 0
abmls$openretire[!is.na(abmls$Q52)|!is.na(abmls$Q51)] <- 1
table(abmls$openretire)
abmls$openproud <- 0
abmls$openproud[!is.na(abmls$Q49)|!is.na(abmls$Q52)] <- 1
table(abmls$openproud)
```

Now, I can test how the *gammas* differ across experimental conditions.

```
group_by(abmls, openretire) %>%
  summarise(
    count = n(),
    mean = mean(topic1, na.rm = TRUE),
    sd = sd(topic1, na.rm = TRUE)
  )
## 'summarise()' ungrouping output (override with '.groups' argument)
## # A tibble: 2 x 4
##   openretire count mean    sd
##         <dbl> <int> <dbl> <dbl>
## 1           0   101 0.443 0.424
## 2           1   100 0.291 0.410
aov <- aov(topic1 ~ openretire, data = abmls)
summary(aov)
##              Df Sum Sq Mean Sq F value Pr(>F)
## openretire    1    1.15  1.1512    6.628 0.0108 *
## Residuals   198   34.39  0.1737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
group_by(abmls, openretire) %>%
  summarise(
    count = n(),
    mean = mean(topic2, na.rm = TRUE),
    sd = sd(topic2, na.rm = TRUE)
  )
## 'summarise()' ungrouping output (override with '.groups' argument)
## # A tibble: 2 x 4
##   openretire count mean    sd
##         <dbl> <int> <dbl> <dbl>
## 1           0   101 0.378 0.400
## 2           1   100 0.338 0.433
aov <- aov(topic2 ~ openretire, data = abmls)
```



```
summary(aov)
##              Df Sum Sq Mean Sq F value Pr(>F)
## openretire    1    0.08 0.07946    0.458  0.499
## Residuals   198   34.36 0.17355
## 1 observation deleted due to missingness
group_by(abmls, openretire) %>%
  summarise(
    count = n(),
    mean = mean(topic3, na.rm = TRUE),
    sd = sd(topic3, na.rm = TRUE)
  )
## 'summarise()' ungrouping output (override with '.groups' argument)
## # A tibble: 2 x 4
##   openretire count mean    sd
##         <dbl> <int> <dbl> <dbl>
## 1           0   101 0.179 0.323
## 2           1   100 0.371 0.437
aov <- aov(topic3 ~ openretire, data = abmls)
summary(aov)
##              Df Sum Sq Mean Sq F value    Pr(>F)
## openretire    1   1.836   1.8355   12.48 0.000512 ***
## Residuals   198  29.125   0.1471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

The results of ANOVA tests show that topic 1 is less often mentioned in the 'retire' condition, while topic 3 is more often mentioned in the 'retire' condition. If we asked respondents to imagine how life would like at the moment they retire (versus 'in 20 years'), they are more likely to name things related to improving society instead of having a steady job and a stable income.

```
group_by(abmls, openproud) %>%
  summarise(
    count = n(),
    mean = mean(topic1, na.rm = TRUE),
    sd = sd(topic1, na.rm = TRUE)
  )
## 'summarise()' ungrouping output (override with '.groups' argument)
## # A tibble: 2 x 4
##   openproud count mean    sd
##         <dbl> <int> <dbl> <dbl>
## 1           0   101 0.365 0.419
## 2           1   100 0.370 0.428
aov <- aov(topic1 ~ openproud, data = abmls)
summary(aov)
##              Df Sum Sq Mean Sq F value Pr(>F)
## openproud    1    0.00 0.00106    0.006  0.939
## Residuals   198   35.54 0.17951
## 1 observation deleted due to missingness
group_by(abmls, openproud) %>%
  summarise(
    count = n(),
    mean = mean(topic2, na.rm = TRUE),
```

```

    sd = sd(topic2, na.rm = TRUE)
  )
## 'summarise()' ungrouping output (override with '.groups' argument)
## # A tibble: 2 x 4
##   openproud count mean    sd
##   <dbl> <int> <dbl> <dbl>
## 1      0    101 0.355 0.413
## 2      1    100 0.362 0.421
aov <- aov(topic2 ~ openproud, data = abmls)
summary(aov)
##               Df Sum Sq Mean Sq F value Pr(>F)
## openproud      1   0.00 0.00211   0.012  0.912
## Residuals    198  34.44 0.17394
## 1 observation deleted due to missingness
group_by(abmls, openproud) %>%
  summarise(
    count = n(),
    mean = mean(topic3, na.rm = TRUE),
    sd = sd(topic3, na.rm = TRUE)
  )
## 'summarise()' ungrouping output (override with '.groups' argument)
## # A tibble: 2 x 4
##   openproud count mean    sd
##   <dbl> <int> <dbl> <dbl>
## 1      0    101 0.280 0.398
## 2      1    100 0.269 0.392
aov <- aov(topic3 ~ openproud, data = abmls)
summary(aov)
##               Df Sum Sq Mean Sq F value Pr(>F)
## openproud      1   0.006 0.00616   0.039  0.843
## Residuals    198 30.954 0.15633
## 1 observation deleted due to missingness

```

Using the words ‘When would you be proud?’ instead of ‘When would you be satisfied?’ does not yield significant differences.

5. Covariates

Finally, I correlate the *gammas* with other survey instruments. In another survey instrument, we asked respondents what they find important in their future job.

```

x <- cbind(abmls$topic1, abmls$topic2, abmls$topic3, abmls$job_1, abmls$job_2, abmls$job_3, abmls$job_4)
x <- na.omit(x)
x <- cor(x)
x <- round(x, 3)
x
##           [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
## [1,]  1.000 -0.558 -0.483  0.107 -0.025 -0.019  0.016 -0.033  0.033
## [2,] -0.558  1.000 -0.457 -0.123  0.035 -0.083 -0.157 -0.154  0.033
## [3,] -0.483 -0.457  1.000  0.015 -0.009  0.107  0.149  0.198 -0.070
## [4,]  0.107 -0.123  0.015  1.000  0.242  0.300  0.324  0.263 -0.050
## [5,] -0.025  0.035 -0.009  0.242  1.000  0.418  0.149 -0.040  0.014

```

```
## [6,] -0.019 -0.083 0.107 0.300 0.418 1.000 0.343 0.229 0.012
## [7,] 0.016 -0.157 0.149 0.324 0.149 0.343 1.000 0.348 0.177
## [8,] -0.033 -0.154 0.198 0.263 -0.040 0.229 0.348 1.000 0.057
## [9,] 0.033 0.033 -0.070 -0.050 0.014 0.012 0.177 0.057 1.000
## [10,] 0.011 0.020 -0.033 0.179 0.175 0.387 0.272 0.247 -0.080
## [11,] 0.212 -0.026 -0.200 0.041 0.116 -0.075 -0.040 -0.025 0.171
##      [,10] [,11]
## [1,] 0.011 0.212
## [2,] 0.020 -0.026
## [3,] -0.033 -0.200
## [4,] 0.179 0.041
## [5,] 0.175 0.116
## [6,] 0.387 -0.075
## [7,] 0.272 -0.040
## [8,] 0.247 -0.025
## [9,] -0.080 0.171
## [10,] 1.000 0.108
## [11,] 0.108 1.000
```

Topic 1 (stable income) correlates .11 with preferring a secure job, -.03 with preferring high income and .21 with preferring an easy work load. Topic 2 (time for your family) correlates -.12 with preferring a secure job, -.16 with interesting job, and -.15 with job that is useful to society. Topic 3 (contributing to society) correlates .2 with preferring a job that is useful to society.

The service motivation scale, which measures the extent to which respondents find it important to contribute to society in their job, correlates -.18 with topic 2 and .26 with topic 3:

```
x <- cbind(abmls$topic1, abmls$topic2, abmls$topic3, abmls$servmot)
x <- na.omit(x)
x <- cor(x)
x <- round(x, 3)
x
##      [,1] [,2] [,3] [,4]
## [1,] 1.000 -0.558 -0.483 -0.063
## [2,] -0.558 1.000 -0.457 -0.184
## [3,] -0.483 -0.457 1.000 0.262
## [4,] -0.063 -0.184 0.262 1.000
```

These results provide evidence on the external validity of the topics and confirm the interpretation based on the word-topic probabilities. Topic 1 is associated with a secure income, although not necessarily a high income. Topic 2 is associated with family values and a slight disinterest with the broader society. Topic 3 is the most prosocial one, in which contributing to society is deemed important.

The *gamma* scores can be used in further (regression) analyses. An issue is their U-shaped distribution, with many values close to 0 and close to 1. However, they are interval variables and no score is exactly 0 or exactly 1. Re-running the correlations with the natural logarithms of the *gamma*s yielded similar results.

6. Concluding remarks

For topic modeling (as for all text analysis) you need sufficient data. We explicitly asked respondents to elaborate on their answers, and this may have helped to get descriptions of multiple sentences. Although the sample (n=200) was small, topic modeling seems to work here.

LDA yielded three sensible topics, and their covariates provide evidence of external validity. As such the open answers can be a measure of how university students see their future, which is more flexible than the standard batteries of closed-ended survey items. The method is more inductive, which fits the objectives of the A Broader Mind Longitudinal Survey. It allows researchers to find surprising elements in their data, while respondents may enjoy answering such an open question after dozens of Likert scales.