* 1. **What type of data are the sentiment values i.e., the values {-5, -4, …, 5}?**

While the words themselves are nominal, the values assigned to them appear to be interval data. They are assigned numeric values, but 2 isn’t necessarily twice as good as 1, and -2 isn’t necessarily twice as bad as -1.

Professor answer:

|  |  |
| --- | --- |
|  |  |
|  | |
| **Dropbox Feedback** |  |
| The choices of a type of data are nominal, ordinal, interval, or ratio.  See the Week 2 From the Expert content.  The sentiment values are ordinal data, which means that they can be ordered.  Differences and ratios of sentiment values may not have much meaning.  It requires an interpretative judgement to assess the usefulness of an interval or ratio type of data. | |

* 1. **Compute a mean value for the sentiments.**

****

**where is the number of occurrences of the th sentiment.**

-5 sentiments 99

-4 sentiments 474

-3 sentiments 641

-2 sentiments 1050

-1 sentiments 1549

0 sentiments 0

1 sentiments 1174

2 sentiments 1749

3 sentiments 1376

4 sentiments 276

5 sentiments 2

(-5)(99)+(-4)(474)+(-3)(641)+(-2)(1050)+(-1)(1549)+(0)(0)+(1)(1174)+(2)(1749)+(3)(1376)+(4)(276)+(5)(2)

/

99+474+641+1050+1549+0+1174+1749+1376+276+2

Equals

-495+-1896+-1923+-2100+-1549+1174+3498+4128+1104+10

/

8390

Equals

1951/8390 = 0.2325

I wanted to use R to do my calculations for me. I am very new to R so I am quite inefficient and am sure there are easier ways to do this, but I just wanted to familiarize myself with the basic navigation & functionality of R. This is essentially a bunch of manual arithmetic, there are certinaly more dynamic ways to do this than what I'm about to display. Here is my sentiment analysis from my 100mb tweet output text file:

-5 sentiments  99  
-4 sentiments  474  
-3 sentiments  641  
-2 sentiments  1050  
-1 sentiments  1549  
0 sentiments  0  
1 sentiments  1174  
2 sentiments  1749  
3 sentiments  1376  
4 sentiments  276  
5 sentiments  2

My first goal was to reproduce this in R:

> mack\_value<-c(-5:5)  
> sentiments<-"Sentiments"  
> mack\_counts<-c(99,474,641,1050,1549,0,1174,1749,1376,276,2)

> mack\_df<-data.frame(mack\_value,sentiments,mack\_counts)

So that reproduced my dataset in R. I'm sure I could have just imported a CSV to R, but I need practice. I'll spare you the suspense and the redundancy - if I type "mack\_df" in R, I get output exactly like I have above.

I made a subset of my data frame just to get rid of "Sentiments". It was unnecessary:

> mack\_subset<-subset(mack\_df,sentiments="Sentiments",select=c(mack\_value,mack\_counts))

Now to calculate my numerator:

> mack\_num\_test <- c(mack\_subset$mack\_value\*mack\_subset$mack\_counts)

> mack\_num\_test

 [1]  -495 -1896 -1923 -2100 -1549     0  1174  3498  4128  1104    10

I will sum these values to get my numerator:

> mack\_numerator<-sum(mack\_num\_test)

> mack\_numerator

[1] 1951

Now to calculate my denominator (excuse my misspelling below!):

> mack\_denomanator<-sum(mack\_subset$mack\_counts)

> mack\_denomanator

[1] 8390

Now that I have my numerator & demoninator, I can get my weighted mean:

> mack\_numerator/mack\_denomanator

[1] 0.2325387

* 1. **What, if anything, does this mean value mean?**

It means that the overall tone of the tweets I analyzed is more positive than negative. Although there is no technical answer to this, the number being close to zero makes me think this is not an extreme indication of positivity of the tweets, but rather mild.

Professor feedback:

To calculate a mean implies that the data are ratio data.  One consequence of this assumption is that words with a positive sentiment value of 4 are twice as positive as those with a sentiment value of 2.  Moreover, the words with a sentiment value of 5 are 2 1/2 times more positive as those with a sentiment value of 2.  Also, those with a sentiment value of -1 are equally negative to those with a sentiment value of 1.  And the examples continue.  This may stretch the interpretations of the sentiment values.

* 1. **In calculating the mean, what additional assumptions about the type of the sentiment data have been made?**

I assume the population is a normal distribution. I would think we need equal, if not close, number of positive & negatively associated words. Say the amount of files in the AFINN was only 5 words, and 4 of those were negative (good +2, bad -1, terrible -3, horrible -4, abysmal -5), your odds of getting a negative tone to your tweet are higher than that of a positive tone.

Professor Feedback:

To consider the numerical imbalance in the number of words for each sentiment value, supose that there were 999 words with a sentiment of -1 and one word with a sentiment of 1. Moreover, assume that all words in this 1000 word sentiment dictionary were equally likely to occur in any given sequence of words. (This assumption may not be warranted.) Then you are about 1000 times as likely to encounter a negative word as you are a positive word.

* 1. **Suppose that in AFINN-111.txt that there were different numbers of words that have been rated -5, -4, -3, …, 4, 5. What effect would this have on your analysis? What might be a way do compensate for this problem?**

As stated in my assumptions, if there were inequivalent numbers of words with positive values than negative, or vice versa, this could skew the results. To compensate for this problem I think the amount of words should be equivalent at each range.