Sentiment Analysis of Amazon Reviews

Tam Tran-The and Jordan Browning
STAT495 Advanced Data Analysis

Description of problem

- Amazon currently sells 480 million products
- Popular products can have thousands of reviews
- How can we summarize reviews and make sense of them for business decisions?

Description of project

- Used sentiment analysis to break Amazon reviews down into their polarity (positive/negative) and their emotions (joy, anger, surprise, etc.)
- Can extract the attitudes and opinions of consumers
- Provides insights into consumer preferences and concerns

Our Data

- Data is taken from Julian McCauley, a UCSD professor
- Amazon reviews between May 1996 and July 2014
- Includes the rating, text of review, product metadata
- Chose to focus on categories of clothing, health, and sports (but same methods apply to all categories)

Sentiment Analysis (Opinion Mining)

- Seeks to extract feelings and opinions from text
- Often relies on lists of words that represent each emotion
- Then uses machine learning algorithms to classify text into a sentiment category:
  - Polarity: positive, negative, or neutral
  - Emotional states: happy, sad, angry, etc.

What emotions can you extract from this review?

“The material is too hard, shrank after washing, do not meet the stated features. I do not recommend purchasing this bluejean, sorry I wasted my money on this product and the worst is that nobody is responsible for the poor quality of the product and the deception suffered.”

Emotion Classification

Emotion Category	Number of words
Anger	335
Disgust	70
Fear	195
Joy	553
Sadness	274
Surprise	95
Total	1542

document-term matrix (DTM)

findFreqTerms()

important words

pmatch() with emotions lexicon

Emotion Category	Class probability $\log\left(\frac{\#\text{ words belong to that category}}{1542}\right)$	Conditional probability for each word in a class $\log\left(\frac{1}{\#\text{ words belong to that category}}\right)$
Anger	1.527	5.814
Disgust	3.092	4.248
Fear	2.068	5.273
Joy	1.025	6.315
Sadness	1.728	5.613
Surprise	2.787	4.554

Number of words

Score

score for each class = “scaled" original score + score assigned to each word * number of words in the class

```
> matrix <- create_matrix("The material is too hard, shrank after washing, do not meet the stated features, I do not recommend purchasing this bluejean, sorry I wasted my money on this product
and the worst is that nobody is responsible for the poor quality of the product and deception suffered")
> words <- findFreqTerms(matrix[1,], lowfreq=1)
> words
 [1] "bluejean"    "deception"   "features"    "hard"        "material"    "meet"        "money"
 [8] "nobody"      "poor"        "product"     "purchasing"  "quality"     "recommend"   "responsible"
[15] "shrank"      "sorry"       "stated"      "suffered"    "washing"     "wasted"      "worst"
> emotions <- lexicon[which(lexicon[,2]=="sadness"),]
> matches <- pmatch(words, emotions[,1], nomatch=0)
> matches
 [1]   0   0   0   0   0   0   0   0 203   0   0   0   0   0   0 241   0   0   0   0   0
> score <- 1.728 + 5.613 * sum(matches > 0)
> score
[1] 12.954
```

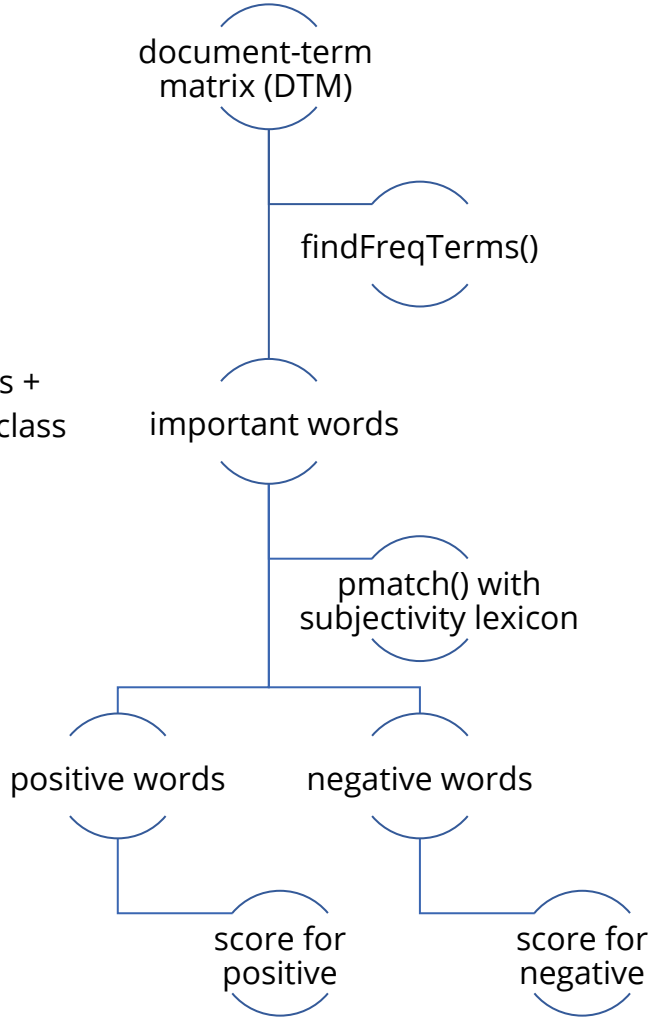
```
> classify_emotion("The material is too hard, shrank after washing, do not meet the stated features, I do not recommend purchasing this bluejean, sorry I wasted my money on this product and the
worst is that nobody is responsible for the poor quality of the product and deception suffered", algorithm="bayes")
     # ANGER # DISGUST # FEAR # JOY # SADNESS # SURPRISE ANGER SCORE        DISGUST SCORE      FEAR SCORE         JOY SCORE          SADNESS SCORE
[1,] "0"     "0"       "0"    "0"   "2"       "0"        "1.46871776464786" "3.09234031207392" "2.06783599555953" "1.02547755260094" "12.9539636605113"
     SURPRISE SCORE     BEST_FIT
[1,] "2.78695866252273" "sadness"
```

Polarity Classification

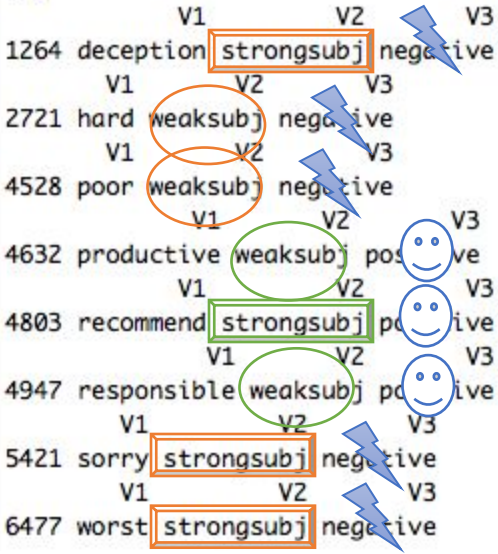
score = “scaled" original score +
score assigned to a weak subj word * number of weak subj words in the class +
score assigned to a strong subj words * number of strong subj words in the class

Polarity Category	Number of words
Positive	2324
Negative	4175
Total	6518

	Class probability $\left\|\log\left(\frac{\#\text{ words belong to that category}}{6518}\right)\right\|$	Conditional prob for a weak subj word $\left\|\log\left(\frac{1}{\#\text{ words belong to that category}}\right)\right\|$	Conditional prob. for a strong subj word $\left\|\log\left(\frac{0.5}{\#\text{ words belong to that category}}\right)\right\|$
+	1.031	7.751	8.444
-	0.445	8.337	9.03



```
> index <- pmatch(words, lexicon[,1], nomatch=0)
> index
 [1]    0 1264    0 2721    0    0    0    0 4528 4632    0    0 4803 4947    0 5421    0    0    0    0 6477
> for (i in 1:21) {
+   if (index[i] != 0) print(lexicon[index[i],])
+ }
```

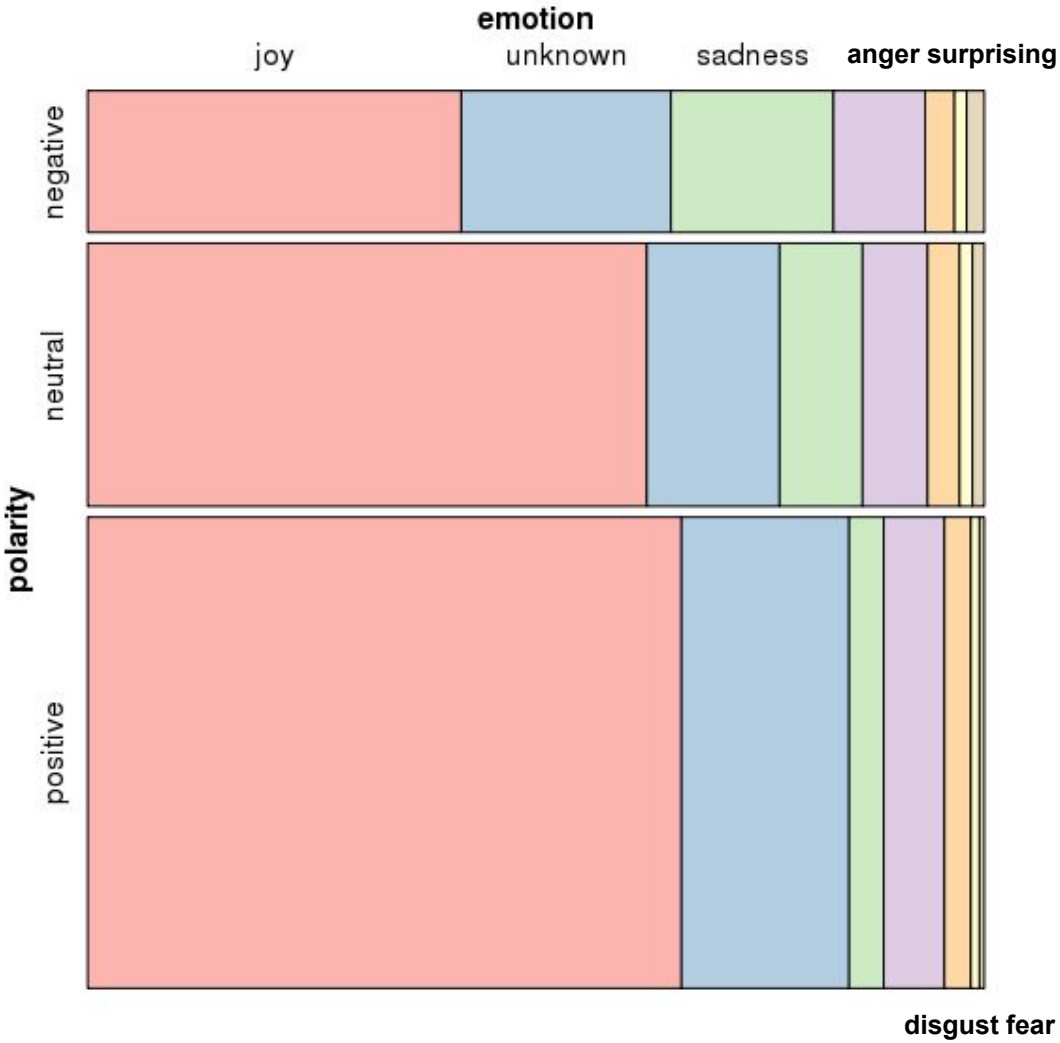


Score for positive category = 1.031 + 2 * 7.751 + 1 * 8.444
= 24.977

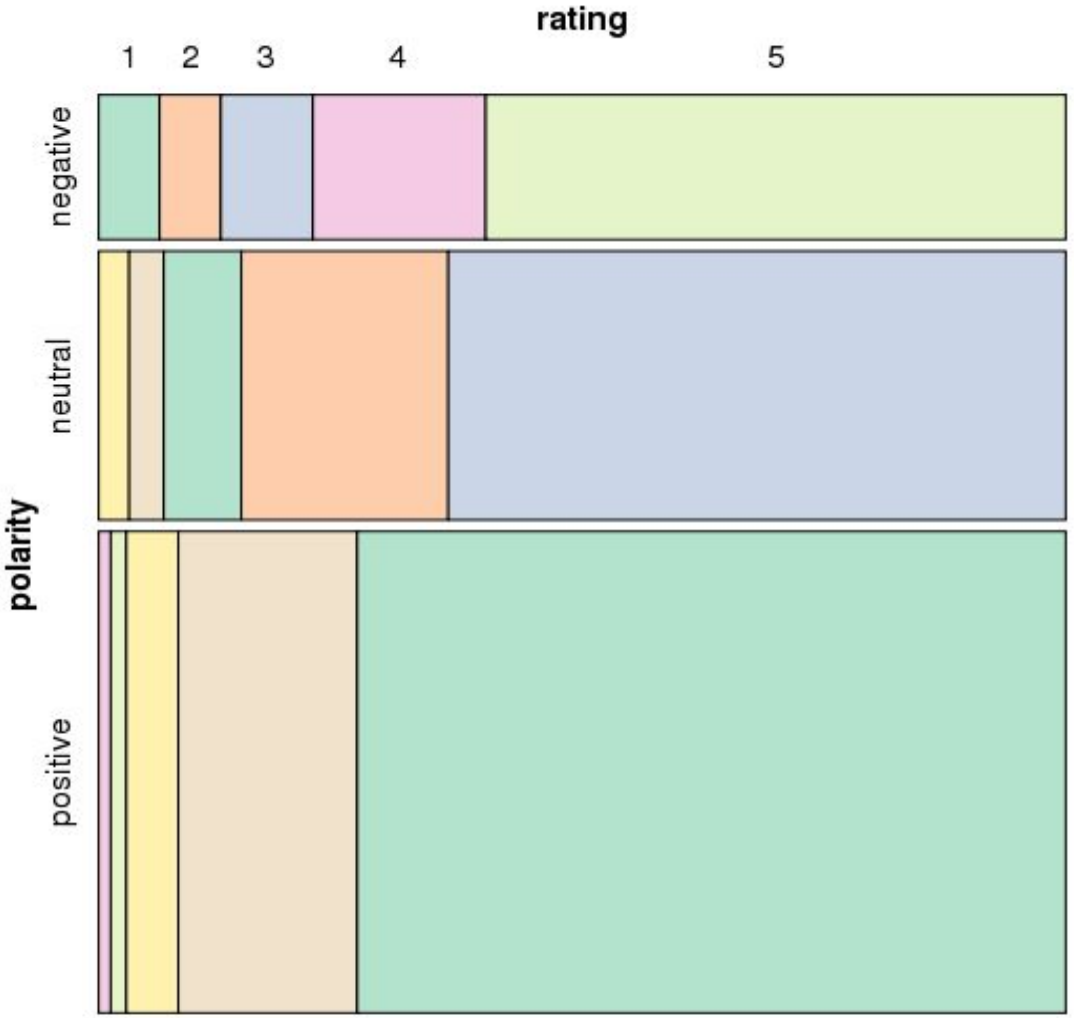
Score for negative category = 0.445 + 2 * 8.337 + 3 * 9.03
= 44.209

```
> classify_polarity("The material is too hard, shrank after washing, do not meet the stated features, I do not recommend purchasing this bluejean, sorry I wasted my money on this product and the
worst is that nobody is responsible for the poor quality of the product and deception suffered", algorithm="bayes")
     #POS #NEG POS SCORE          NEG SCORE          POS/NEG             BEST_FIT
[1,] "3"  "5"  "24.9775602759011" "44.2092429502172" "0.564985025960015" "negative"
```

Overlap Between Emotion and Polarity Classification



Overlap Between Rating and Polarity Classification



Discrepancy

“I had high hopes for this necklace, since all the other cheap ones I bought were absolutely just stunning and perfect. This one, however, is not. For starters, the pic gives the impression that the stones on the tail are all different colors. They aren't. They are all opal looking, and not centered at all. It looks like someone had a child paste these stones on. They are nowhere near center, like in the picture, and just look ridiculous. The only saving thing about this necklace is the peacock body is nice. I'll have to take off the old stones, on the tail, and replace and center them. Other than that, save your money unless you want a piece that you will have to take apart and recreate. Not worth the money, in my opinion.”

Rating: 1

Polarity: positive

Emotion: joy

“I had high hopes for this necklace, since all the other cheap ones I bought were absolutely just stunning and perfect. This one, however, is not. For starters, the pic gives the impression that the stones on the tail are all different colors. They aren't. They are all opal looking, and not centered at all. It looks like someone had a child paste these stones on. They are nowhere near center, like in the picture, and just look ridiculous. The only saving thing about this necklace is the peacock body is nice. I'll have to take off the old stones, on the tail, and replace and center them. Other than that, save your money unless you want a piece that you will have to take apart and recreate. Not worth the money, in my opinion.”

```
#ANGRY: 1 #DISGUST: 0 #FEAR: 0 #JOY: 4 #SADNESS: 0 #SURPRISE: 0
 ANGER SCORE: 7.341 DISGUST SCORE: 3.092 FEAR SCORE: 2.068 JOY SCORE: 26.287 SADNESS SCORE: 1.728 SURPRISE SCORE: 2.
787
 BEST FIT: joy
```

```
#POS: 11 #NEG: 5
 POS SCORE: 93.224 NEG SCORE: 44.902
 POS/NEG: 2.076 BEST FIT: positive
```

```
positive joy
```

Discrepancy

“I'm allergic to scented detergents (both hives and sore throat and congestion) and this is great for that. It gets clothes clean and I don't have any problem with my allergies in regards to this soap. I recommend it over other similar products only if you can get it when it's on sale.”

Rating: 5

Polarity: negative

Emotion: anger

“I'm allergic to scented detergents (both hives and sore throat and congestion) and this is great for that. It gets clothes clean and I don't have any problem with my allergies in regards to this soap. I recommend it over other similar products only if you can get when it's on sale.”

```
#ANGRY: 1 #DISGUST: 0 #FEAR: 0 #JOY: 1 #SADNESS: 1 #SURPRISE: 1
 ANGER SCORE: 7.341 DISGUST SCORE: 3.092 FEAR SCORE: 2.068 JOY SCORE: 7.341 SADNESS SCORE: 7.341 SURPRISE SCORE: 7.3
41
 BEST FIT: anger
```

```
#POS: 3 #NEG: 5
 POS SCORE: 25.671 NEG SCORE: 43.516
 POS/NEG: 0.59 BEST FIT: negative
```

```
negative anger
```

Shiny App

https://r.amherst.edu/apps/m5ttranthe/Amazon%20sentiment%20analysis/

Implications

- With so much data available, businesses should use sentiment analysis to gain a better understanding of customers
- Potential uses
  - Find product features are frequently associated with negativity to improve product
  - Improve marketing by emphasizing what customers like about a product
  - Detect spam reviews intended to promote or demote a product

Future Directions

- Could expand shiny app to analyze reviews of any product
  - Requires the direct web scraping
- Using other methods of sentiment analysis to improve accuracy