Social Media Moderation with Sentiment Analysis

Anthony Buchanan

Foundations of Data Science

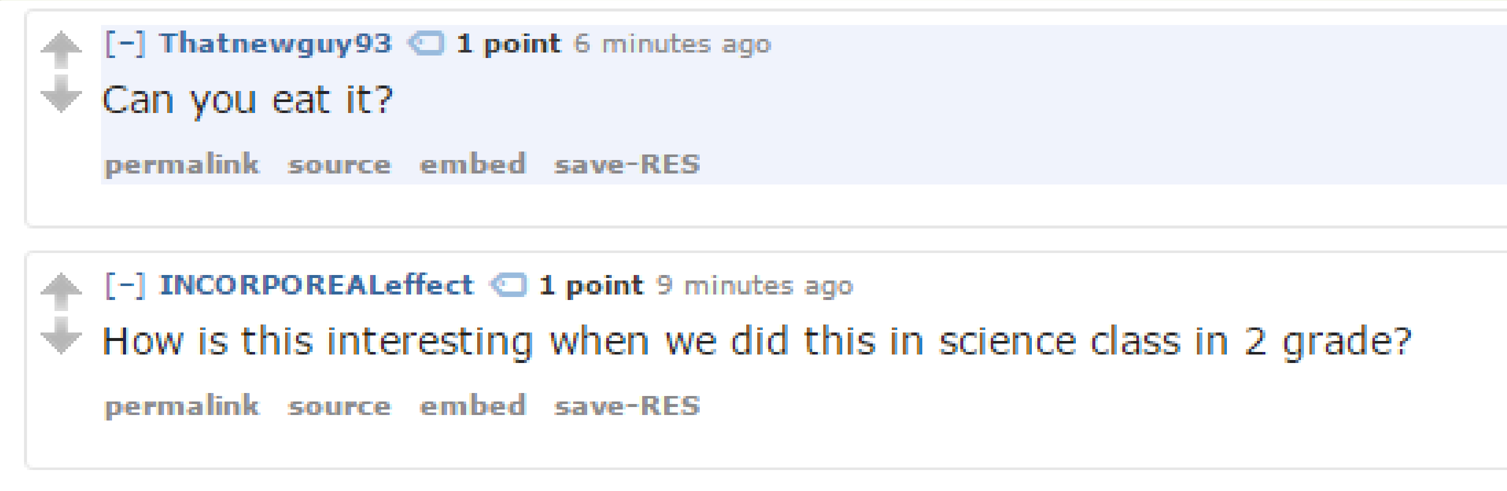
August 2016

The Problem: Something's Wrong on the Internet

- Reddit – "The front page of the internet", a popular online community, with over 230 million unique monthly visitors
- Users post and vote on comments in topic based communities called subreddits
- Comments and users may be popular but still "generally make Reddit worse for everyone else"[1]
- Currently way to identify comments based that contribute to the community based on their overall sentiment

[1] Reddit finally bans its most infamous racist communities because they 'made recruiting here more difficult', Business Insider, Matt Weinberger, 5 Aug 2015, http://www.businessinsider.com/reddit-bans-coontown-2015-8

Reddit Comments



[–] Thatnewguy93 1 point 6 minutes ago

Can you eat it?

permalink source embed save-RES

[–] INCORPOREALeffect 1 point 9 minutes ago

How is this interesting when we did this in science class in 2 grade?

permalink source embed save-RES

- Do these comments contribute positively? Negatively? Or make no meaningful contribution?
- What is the sentiment or tone of these comments?

The Project: Sentiment Analysis by Classification Model

- Use Naïve Bayes classification model to sort comments into "Positive" (1), "Neutral" (0) and "Negative" (-1)
- Train the model with a sentiment analysis lexicon of 6,800 positive and negative English words[2]
- A comment's score is the sum of the score of its constituent words after standardizing text

[2] "Opinion Mining, Sentiment Analysis, and Opinion Spam Detection, Hu and Liu, https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

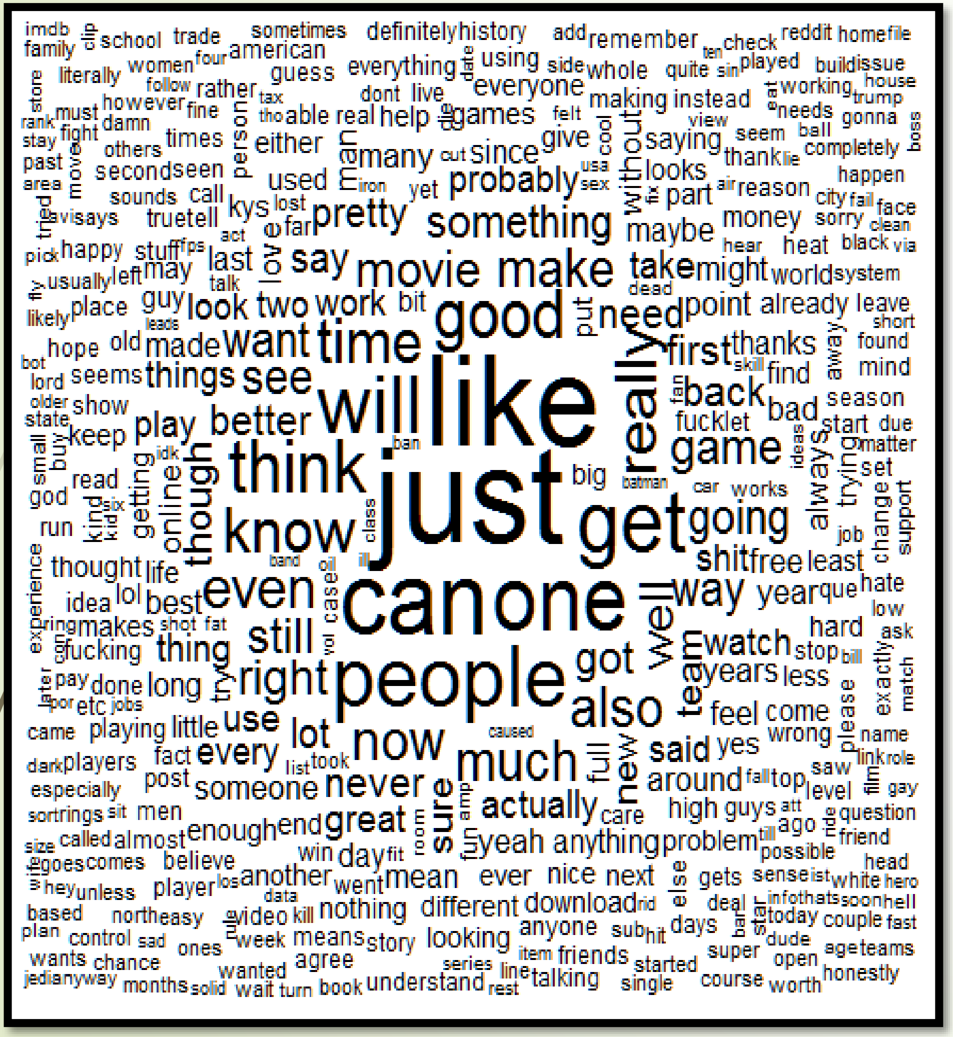
The Data: You Get an API…

- Data was collected from the Reddit API (real time stream)
  - Over one million comments from over 17,000 subreddits
  - Data was collected over a four day period (5 July to 8 July 2016)
  - Id, subreddit, author, votes, comment text, and created date
- Data was scraped using libraries such as rCurl and jsonlite and saved to a database using r.utils and DBI libraries
- Data was cleaned to remove characters, duplicate comments, bot/moderator comments and more

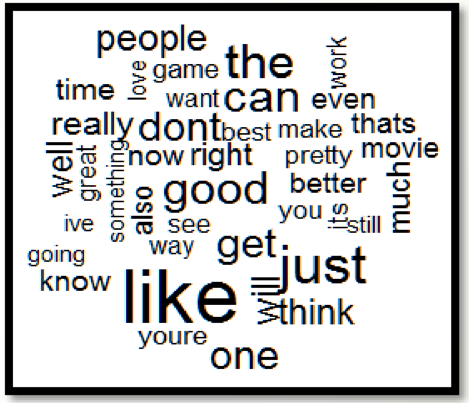
The Data: Developing the Training Set

- A set of 30,000 comments were selected with 75% placed into the training set and 25% placed in the test set
- An r script was run on the test set using the 2 positive and negative opinion lexicons to score the sentences
  - Words were evaluated individually and then summed to create the overall comment score
- Scores ranged from -55 to 47 but most comments clustered around a score of -1, 0, or 1. Many had a score of 0

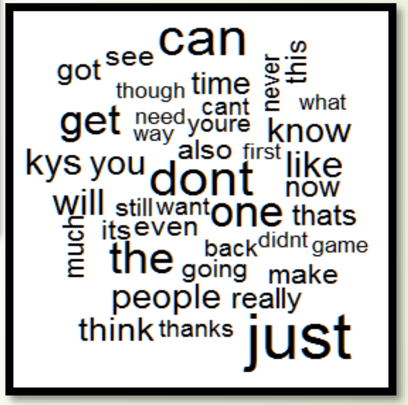
The Data: Developing the Training Set (2)



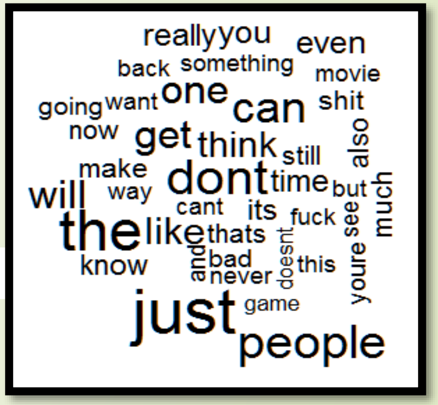
Training Data (All)



Training Data (Positive)

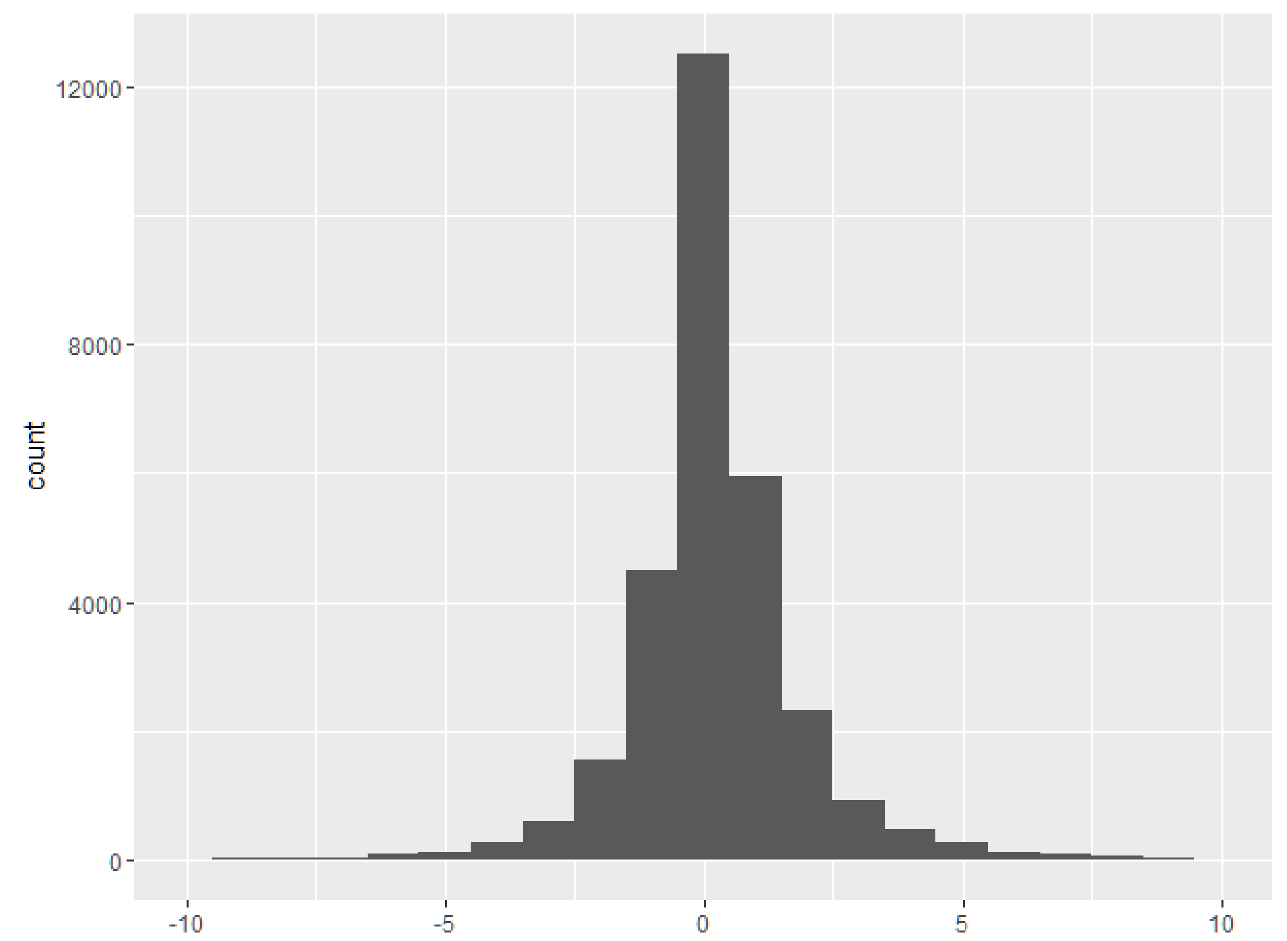


Training Data (Neutral)

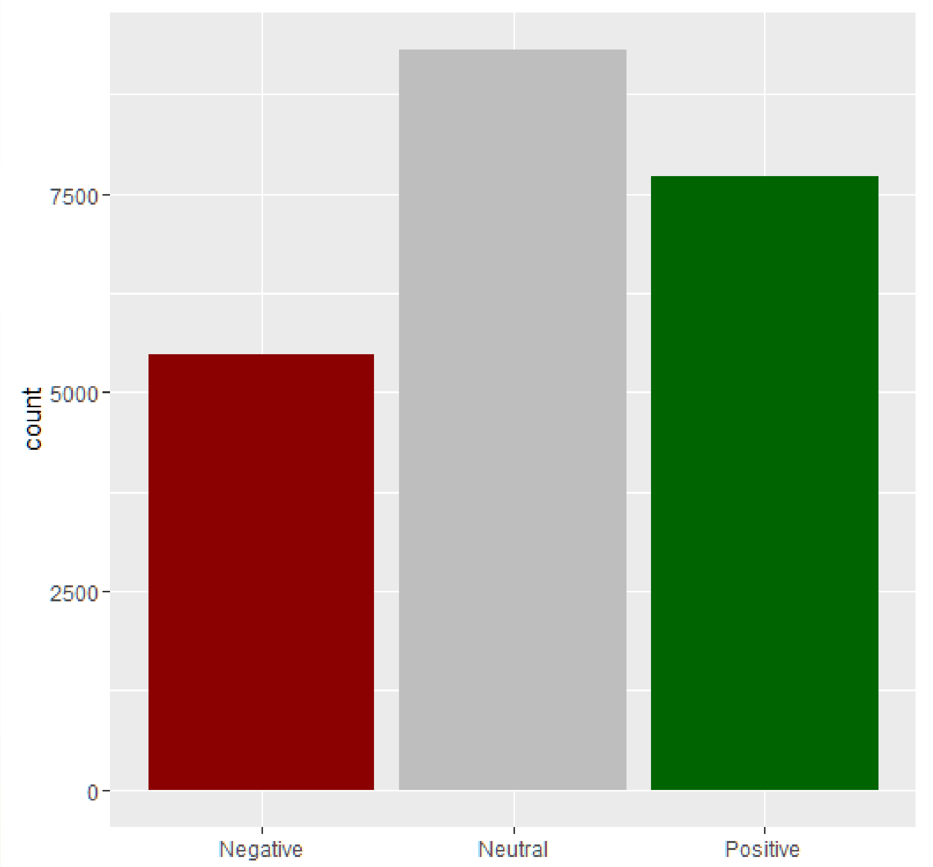


Training Data (Negative)

The Data: Developing the Training Set (3)



Sentiment vs Count

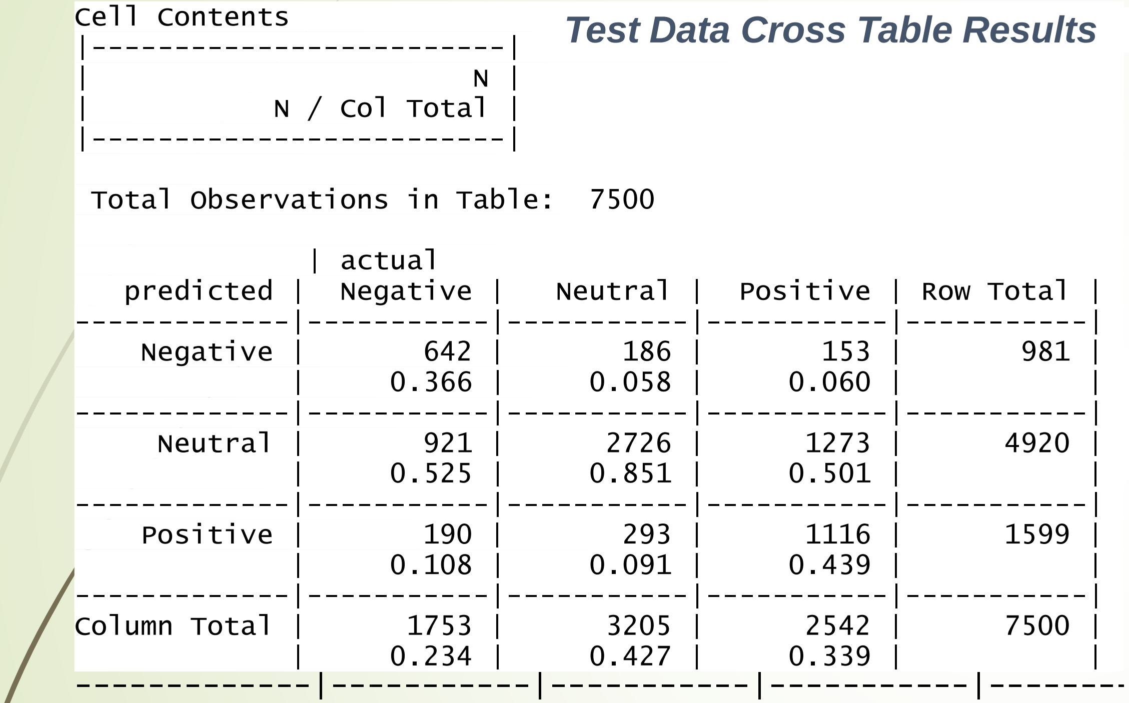


Sentiment Categories vs Count

The Analysis: Developing the Model

- Using the text mining library tm the data was cleaned further and a corpus and document term matrix was created from the comment test and training sets
- The library e1071 was used to create the naïve bayes model
- Model performance was evaluated with gmodels library and pROC library
- The model correctly identifies 85% of neutral comments, 44% of positive comments and 37% of negative comments
- Over 50% of both the incorrectly classified positive and negative comments were classified as neutral

The Analysis: Developing the Model (2)

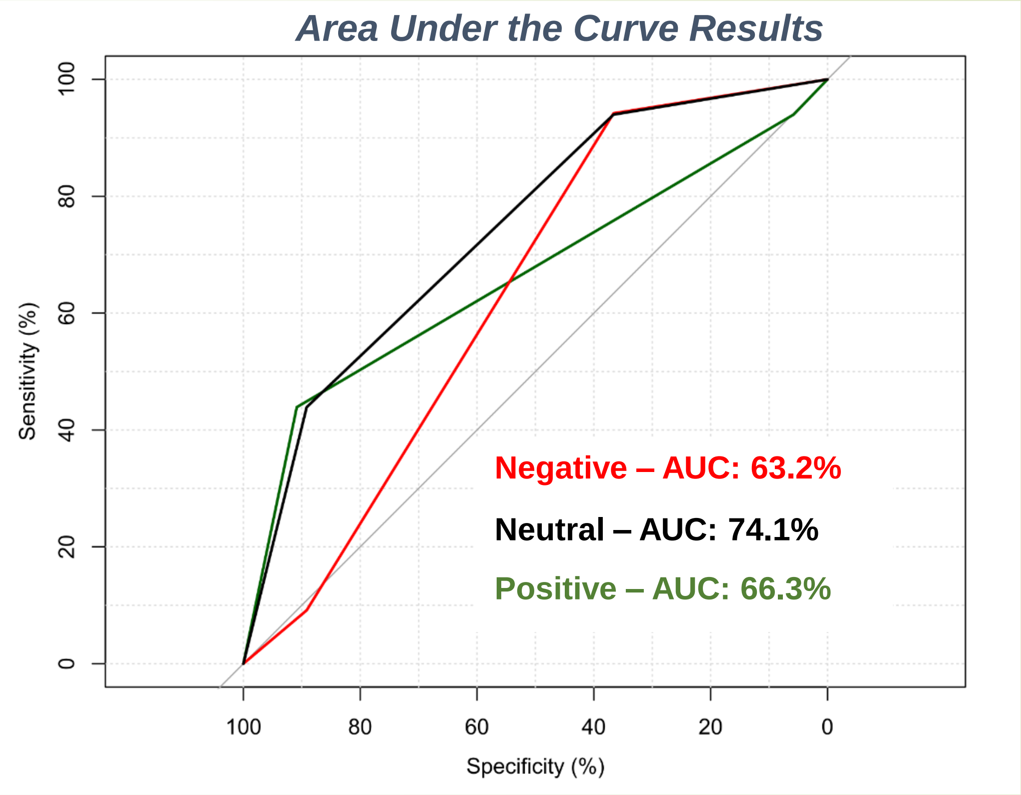


Test Data Cross Table Results

```
Cell Contents
|-------------------------|
|                       N |
|           N / Col Total |
|-------------------------|

 Total Observations in Table:  7500

                | actual
   predicted | Negative |    Neutral |   Positive | Row Total |
-------------|------------|------------|------------|------------|
    Negative |        642 |        186 |        153 |        981 |
             |      0.366 |      0.058 |      0.060 |            |
-------------|------------|------------|------------|------------|
     Neutral |        921 |       2726 |       1273 |       4920 |
             |      0.525 |      0.851 |      0.501 |            |
-------------|------------|------------|------------|------------|
    Positive |        190 |        293 |       1116 |       1599 |
             |      0.108 |      0.091 |      0.439 |            |
-------------|------------|------------|------------|------------|
Column Total |       1753 |       3205 |       2542 |       7500 |
             |      0.234 |      0.427 |      0.339 |            |
-------------|------------|------------|------------|--------
```



Area Under the Curve Results

			Negative		Positive		Neutral	
True Positive	False Negative		642	1111	1116	1426	2726	479
False Positive	True Negative		339	5408	483	4475	2194	2101
	F1 Score		47.0%		53.9%		67.1%	

True and False Negatives and Positives

The Results: Sentiment Analysis

- Although the model could use improvement, it still provided a classification that could be used to flag users for human review based on consistent negative or positive comments
- Mood badges could provide a fun metric to users
- Model would benefit from a better training set. Data could be classified by humans using micro workers such as Amazon Mechanical Turk
- A more advanced classification model such as a neural net could be explored in the future

Thank You!