Text Mining and Pattern Recognition for Online Reviews

Maudeline Deus[1], Reinaldo Sanchez-Arias, Ph.D[2]
[1]Miami Dade College, Wolfson Campus, Miami FL
[2]St. Thomas University, School of Science, Miami Gardens, FL

Introduction

Thousands of social media posts, biographies, and comments can be interpreted using sentiment analysis. Combined with digitalized text with a process called text mining, it is easier to get background information on a data set and to proceed with a thorough approach. Sentiment analysis is a tool used in modern day technology to decipher what would normally be considered a negative outreach from what would be a positive one. In this work, three different sentiment lexicons, based on unigrams (i.e. single words) were used: NRC, Afinn, and Bing [3]. Each lexicon is derived from a single English word and are assigned different scores of positive/negative sentiments. Each lexicon scores a different way from the others. Affin scores a word with a number, which may range from -5 to +5. Bing scores a word as either positive or negative. NRC categorizes a word under sentiment type categories such as positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Some lexicons may be more descriptive to the human mind than other lexicons, the information provided by these lexicons helps to organize information from the outside world into a tidy text that computers can generate. This makes the information taken from the dataset more manageable for clustering analysis and other pattern recognition methods.

Data Collection

The case study for the data analysis of online reviews involved web scraping of comments on RateMyProfessors.com (RMP) [4], a review site that allows college and university students to assign ratings to professors and campuses of American, Canadian, and United Kingdom institutions. Users have added more than 19 million ratings, 1.7 million professors and over 7,500 schools. The study focused on the reviews left to mathematics and statistics instructors at Miami Dade College (MDC). Comments for a total of 559 instructors were collected which included every single review along with their tags. As of date, RMP allows the user leaving a comment to add up to 3 different tags when reviewing a professor from a total of 20 different choices (see figure 3). The data used was cleaned and anonymized to not include the name of the instructors. Data tidying, the process of cleaning and structuring datasets to facilitate analysis, was required to account for human errors and incorrect format. Some of the issues that were taken into account and resolved while building the dataset include: incorrect input of course numbers, instructors listed multiple times or in incorrect departments, and instructors that teach in multiple subject areas (e.g. Biology instructors that have also taught introductory level math courses), among others.

```r
stat_2023_sentiment <- tidy_comments %>%
  ungroup() %>%
  filter(course == "STA2023") %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
```

Fig 1: Sample R code snippet used to collect comments left in RMP for reviews of STA2023 Applied Statistics, studied using the Bing sentiment lexicon. Data transformation, filtering, and cleaning was required to tidy up the comments data set for text mining

Data Science with RStudio

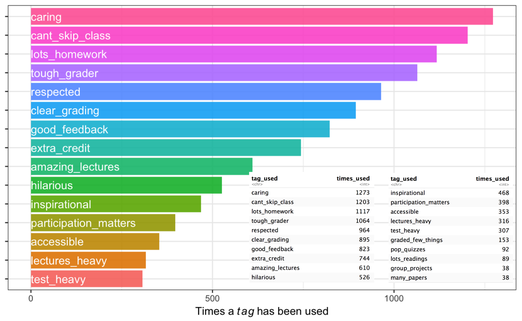
Programming was essential in the development of this project. R [1] is an open source language widely used in the data science community, with focus on statistical data analysis, data visualization and machine learning methods. During this project, tools from the `tidyverse` package [2] were used for data wrangling and data visualization with the help of RStudio, an open source integrated development environment (IDE) for R. In particular we used the `ggplot2` package for *data visualization*, `dplyr` and `stringr` for *data transformation* and summaries, and `rvest` for *web scraping*. Additionally we used tools from the `tidytext` package [3] for word processing and sentiment analysis of the different comments left on reviews.

Fig 2: RStudio logo. RStudio makes R easier to use. It includes a code editor, debugging and visualization tools. Logos for the `ggplot2`, `dplyr`, and `stringr` packages.

Sentiment Analysis and Clustering

Exploratory Data Analysis

Basic exploratory data analysis (EDA) was performed to better understand the distribution of words used in the different comments, as well as to study the tags associated to each review extracted from the RMP website.



tag_used	times_used	tag_used	times_used
caring	1273	inspirational	468
cant_skip_class	1203	participation_matters	398
lots_homework	1117	accessible	353
tough_grader	1064	lectures_heavy	316
respected	964	test_heavy	307
clear_grading	895	graded_few_things	153
good_feedback	823	pop_quizzes	92
extra_credit	744	lots_readings	89
amazing_lectures	610	group_projects	38
hilarious	526	many_papers	38

Fig 3: Frequency of tags used in RMP as part of a review of a mathematics/statistics course (the top 15 tags are shown)

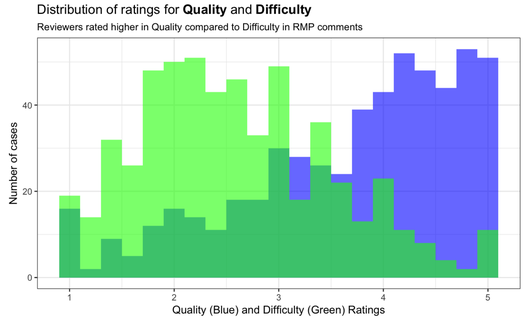


Fig 4: Histogram showing the distribution of ratings for "Quality" and "Difficulty" of mathematics and statistics instructors.

What is Sentiment Analysis?

Sentiment analysis can be thought of as the exercise of taking a sentence, paragraph, document, or any piece of natural language, and determining whether that text's emotional tone is positive, negative or neutral. One way to analyze the sentiment of a text is to consider the text as a combination of its individual words and the sentiment content of the whole text as the sum of the sentiment content of the individual words.

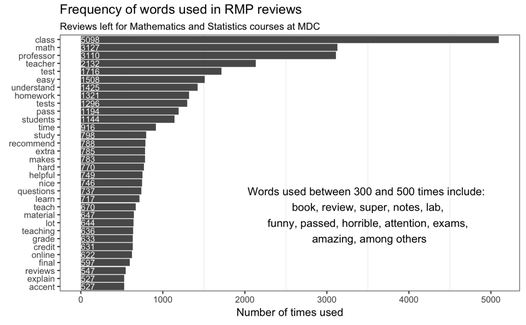


Fig 5: Bar plot showing the most common words used in RMP reviews for math and stats.

Possible emotions like joy, anger, sadness, and so forth can be assigned to a term depending on the specification of the lexicon. Having data in a tidy format, sentiment analysis can be done as an *inner join*, with a database of words and score per lexicon (see Figure 1). Similarly, removing stop words is an *antijoin* operation with a dataset of stop words (e.g. "the", "of", "to", and so forth in English).

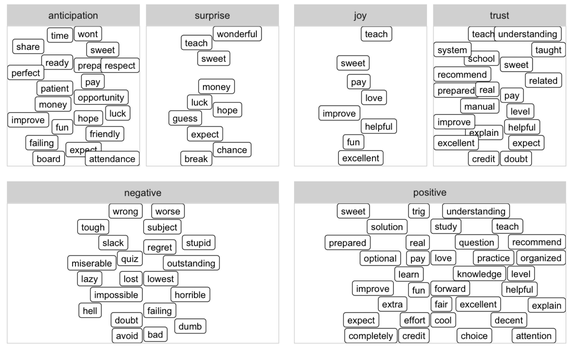


Fig 6: Words used in RMP reviews for MAC 2311 Calculus I, classified by NRC sentiment

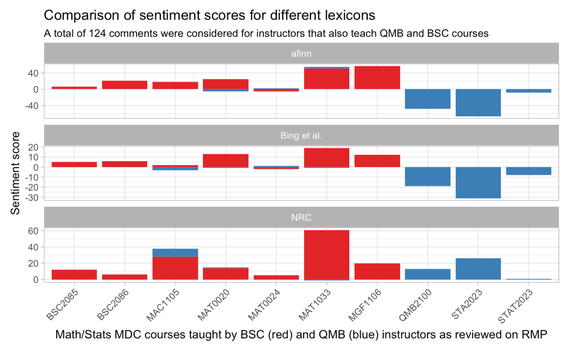


Fig 7: Courses taught by QMB (Business Statistics) instructors showed more negative sentiments than those taught by BSC (Biology) instructors in the dataset considered in this study.

Clustering

Clustering methods deal with finding similarities and structure in a collection of data points. We used k-means clustering [5], a distance-based method, that consists of creating clusters so that the total intra-cluster variation is minimized. In k-means clustering, each cluster is represented by its center (i.e, centroid) which corresponds to the mean of points assigned to the cluster.

The standard implementation is the Hartigan Wong algorithm, which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid:

$$W(C_j) = \sum_{x_i \in C_j} (x_i - \mu_j)^2$$

where x_i is a data point belonging to the cluster C_j , and μ_j is the *mean value* of the points assigned to cluster C_j

The basic steps of the algorithm are outlined below:

1. k centerpoints are randomly initialized.
2. Assign objects to their closest cluster center according to the Euclidean distance.
3. Calculate the centroid or mean of all objects in each cluster.
4. Repeat steps 2 and 3 until no variations in assignments is observed.

k-means

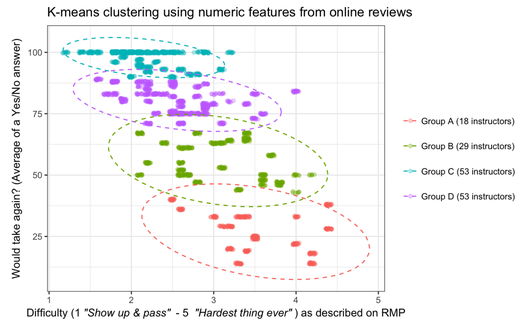


Fig 8: K-means Clustering using the meta-data "Would Take Again?" and "Difficulty" for instructors considered in the data set. (Not all instructors had records for the "Would take again?" question, therefore a smaller amount of data points were considered)

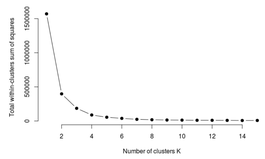


Fig 9: "Elbow method" to determine the number of cluster to consider (i.e. value of k). The total within-cluster sum of square measures the compactness of the clustering and we want it to be as small as possible.

```r
set.seed(123)
# function to compute total within-cluster sum of square
wss <- function(k) {
  kmeans(read_comments[,c("difficulty", "would_take_again")],
         k, nstart = 10 )$tot.withinss
}
# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15
# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)
plot(k.values, wss_values,
       type="b", pch = 19, frame = FALSE,
       xlab= "Number of clusters K",
       ylab= "Total within-clusters sum of squares")
```

Fig 10: R code snippet showing the implementation of the "elbow method" to find a good value of k in k-means clustering

Conclusions

On the website ratemyprofessors.com, different types of reviews can be found. When it comes to mathematics courses generally more negative comments are expected rather than positive. During the analysis, it was found that the tag *"caring"* was the most used tag for mathematics instructors while *"test heavy"* was the least used tag by students to rate their instructors. In the sentiment analysis implementation, it was found that the introductory courses had higher sentiment values than what would be considered to be the higher level courses. While looking at mathematics instructors that also teach BSC and QMB courses, the majority of the courses taught by QMB faculty had a negative sentiment score compared to those that teach BSC courses.

References

[1] R: free software environment for statistical computing and graphics. https://www.r-project.org/
[2] `tidyverse`: an opinionated collection of R packages designed for data science. https://www.tidyverse.org/
[3] `tidytext`: Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools. https://CRAN.R-project.org/package=tidytext
[4] RMP: Rate My Professors http://www.ratemyprofessors.com/About.jsp
[5] James, G. (2017). An introduction to statistical learning: With applications in R. New York: Springer.

Acknowledgements

This work was made possible by the support of the School of Science at St. Thomas University (STU). We also want to highlight the support of our research partners Eliana Espinosa and Sierra Hawthorne, mathematics and data science students at STU, for always being present when needed, and the Faculty and Staff at the School of Science at STU. This project was supported, in part, by U.S. Department of Education grant award P03C1160161 (STEM SPACE). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the respective funding agency.