Using Natural Language Processing (NLP) to Implement Sentiment Analysis and Keyword Extraction on Yale Course Evaluations

Alexander J. Abinader
Advisor: Prof. Robert Wooster

February 2024

1 Background

Natural language processing (NLP) is a modern technology that allows computers to interpret and manipulate human language. The NLP field has experienced explosive growth in the past several years, mainly due to its vast number of important applications. Popular technologies like speech recognition, email filtering, chatbots, predictive text, and many others rely heavily on NLP techniques, making it a highly sought after skill for programmers and data scientists worldwide.

Two of the most popular NLP areas include sentiment analysis and keyword extraction. Sentiment analysis is an extremely valuable procedure that allows computers to evaluate written text through a human lens. For example, there exist many pretrained sentiment analysis models that have been used for a variety of popular applications such as evaluating Amazon product reviews or rating the sentiment of social media posts such as those on Twitter (now X).

Keyword extraction is a separate technique that automatically identifies the key words that summarize the main topics of a written text. A similar process called topic modeling also seeks to identify the key terminology present in text but goes one step further to identify what underlying themes these words represent. Generally, in the context of shorter written pieces (a few words to a short paragraph), keyword extraction proves more effective at identifying terms of interest.

2 Introduction

At the end of each semester, Yale students are given optional surveys to complete where they review their classes for that term. Within these surveys are sets of questions that provide five multiple choice responses. For example, one of the most important questions is "What is your overall assessment of this course?" to which students can reply with 'poor', 'fair', 'good', 'very good', or 'excellent'. Notably, this type of question and its discrete set of responses can be converted into a numerical score. This facilitates the process of creating summary statistics for course evaluations using this quantitative data. A student created, student run website named CourseTable does just that.

CourseTable is a course review website that serves the Yale community by providing information about Yale classes. In addition to official registrar data like class enrollment size, professor, location, etc., CourseTable also uses the official end of semester surveys that students fill out in order to publish a host of descriptive statistics and reviews for Yale's thousands of courses. Namely, CourseTable uses the responses from the multiple choice survey questions about course overview, intellectual challenge, workload, and others in order to create three quantitative ratings: Average Course Rating, Average Workload Rating, and Average Professor Rating.

In addition to multiple choice survey questions, Yale's official end of semester surveys also provide room for written responses, free from any constraints. This data is presented on CourseTable in its raw form. No analysis is performed on these written responses. CourseTable simply lists the questions that were asked in the registrar survey and then prints the entire set of student responses. Analyzing this textual data is just the task for NLP.

3 Project Description

This project's ultimate goal is to use NLP techniques to analyze, interpret, and evaluate written course evaluations by Yale students. Currently, the official Yale course evaluations coupled with CourseTable's additional calculations paint a fairly impressive picture of individual Yale courses. However, I believe that using NLP techniques to analyze written responses can make this data even more informative and accessible. For example, consider large lecture classes that receive hundreds of written reviews. Using NLP, we can summarize the key information so that students can more efficiently search through the thousands of classes offered each semester. Moreover, professors can use this information to get an overview of their students' responses before diving into the individual reviews.

As mentioned in the Introduction, there are two broad NLP techniques we can use to extract meaning from these course evaluations. The first is sentiment analysis. One key question from the surveys is "Would you recommend this course to another stu-

dent? Please explain." Using sentiment analysis, we can hopefully analyze these replies and then calculate a numerical score to summarize student sentiment with respect to the class. The other prompts presented on CourseTable are "What knowledge, skills, and insights did you develop by taking this course?" and "What are the strengths and weaknesses of this course and how could it be improved?" Some form of keyword extraction can be used to tackle both of these, hopefully highlighting the skills, strengths, weaknesses, and areas of improvement for a course without having to read through all of the responses.

In this project, I will set out a series of baseline goals followed by several stretch goals that I may tackle if time permits. First, I will learn sentiment analysis and apply this technique to the recommendation question. To do so, I will use both pretrained models and train models from scratch after manually labeling the dataset in order to run, test, and compare a variety of sentiment analysis procedures for the task at hand. This will involve using rule based models like VADER, machine learning models like random forest, logistic regression, and support vector machines, and deep learning with neural networks.

Next, I will use keyword extraction for the skills prompt in order to summarize student responses. This will be similar in that I'll use both pretrained keyword extraction models and train models from scratch using manually labeled data. As a stretch goal, I will also try this for the strengths/weaknesses/improvements prompt although this has additional complexity given that the model would have to classify terms into these categories.

This project will involve in depth theoretical and practical research of two NLP techniques: sentiment analysis and keyword extraction. The process will include data collection, data cleaning, programming, model selection and training, model evaluation, and comparative analysis. This will culminate in a written report that details each method used, its mathematical theory, the resulting model, and an evaluation of said model. Another stretch goal will be to implement a website/dashboard that can be queried to display this information to a user, perhaps using the best model from this project.

4 Timeline

The following timeline provides a rough estimate of which project elements will be addressed each week:

- Week 4, Feb. 5 - Proposal Submission and Data Collection
- Week 5, Feb. 12 - Data Cleaning and Literature Review
- Week 6, Feb. 19 - Literature Review and Sentiment Analysis Models

- Week 7, Feb. 26 - More Sentiment Analysis Models
- Week 8, Mar. 4 - Sentiment Analysis Evaluation and Wrap Up
- Week 9, Mar. 11 - Keyword Extraction Models
- Week 10, Mar. 18 - More Keyword Extraction Models
- Week 11, Mar. 25 - Keyword Extraction Evaluation and Wrap Up
- Week 12, Apr. 1 - Fine-tuning and Comparative Analysis
- Week 13, Apr. 8 - Theoretical Research and Written Report
- Week 14, Apr. 15 - Final Written Report
- Week 15, Apr. 22 - Buffer for Catch Up/Stretch Goals
- Week 16, Apr. 29 - Buffer for Catch Up/Stretch Goals

5 Deliverables

1. Source code for collecting and cleaning the relevant data
2. Several sentiment analysis models including both pretrained and newly trained
3. Several keyword extraction models including both pretrained and newly trained
4. A written research report that details the NLP techniques used in theory and practical implementation
5. (Stretch) More complex keyword extraction for strengths, weaknesses, and areas of improvement
6. (Stretch) A website or dashboard that allows users to access the NLP summary information