Final Project Part 1: Proposal

Doug Perez

Luddy School of Informatics, Computing, and Engineering

FA22-BL-DSCI-D590-11691: Natural Language Processing

Dr. Olga Scrivner

October 9, 2022

Project Title

Sentiment Analysis for Professional Evaluations

Team

I will be working by myself on a team of one.

Project Description

Objectives

I work as a technical trainer for Amazon Web Services. As a large tech company, we strive to use data to make business decisions. As a result, my work is measured in several quantifiable ways, I am held to certain standards in the metrics collected, and my end-of-year performance will be evaluated in part according to my performance in given metrics.

The primary vehicle that collects data used to measure my performance is individual student satisfaction surveys. These surveys collect both quantitative measurements on a scale and qualitative user comments. Recently, we have started using sentiment analysis to quantify the user comments; however, the results are not perfect. For example, I recently had a student reply “Nothing at all,” to the question, “What would you recommend changing about this course.” A person reading these would interpret the sentiment of the comment as positive; the sentiment analysis algorithm we have implemented interpreted it as negative.

My objective for the final project is to improve the results of our existing sentiment analysis system. Specifically, I would like to build a better sentiment analysis tool for professional user academic feedback. This goal is slightly different than a standard sentiment analysis project in that the audience will largely consist of professionals and the focus of the sentiment will be the delivery of academic information.

Usefulness

If successful, the results of this application would allow us to quantify the qualitative sentiments expressed in literally thousands of student satisfaction surveys every week. Beyond its utility at my company, higher and professional education are universal fields that could improve given the appropriate data with which to make better decisions. While there are obviously a lot of sentiment analysis applications in the world, I think the focus on professional respondants offering sentiments on academic content will satisfy a niche where existing applications are not as clearly targeted. Further, I will be conducting the analysis in 2 languages, which will be discussed in more detail in the next section.

Data

Clearly, the internal customer satisfaction data at AWS is confidential, so I will not be able to build my application using the data set that set my interest in this topic in motion. However, I found a public, commonly used data set that I believe will meet my requirements while also adding a layer to make the results more universally applicable.

The data set I have chosen is the Paper Reviews Data Set from UC-Irvine, where I happen to be a graduate faculty advisor. The data is unprocessed and in JSON format. Here is the URL for the dataset:

https://archive.ics.uci.edu/ml/datasets/Paper+Reviews

The dataset consists of reviews of academic papers. The majority of the reviews are in Spanish with the rest in English. As I am bilingual, I will be able to work in both languages. Further, the bilingual nature of the data supports my initial goal because my company conducts classes all over the world and I am positioning myself to delivery training in both Spanish and English.

I believe this dataset will help me achieve the goal of better classifying my company's student evaluations for two key reasons. First, the audience giving the reviews is academic. Second, the content they are reviewing is also academic. These two points together satisfy my desire to better target the sentiment analysis application beyond areas like product reviews or social media comments. Third, the data is labeled on a 5-point scale from -2 (very negative) to +2 (very positive). This scale aligns nicely with the quantitative evaluations we do at AWS, the majority of which are also on a 5-point scale.

The data contains 405 evaluations and 10 columns. The columns are as follows:

Column Name	Data Type
Timespan	Datetime
Paper ID	Integer
Preliminary decision	Label
Review ID	Integer
Text	Text
Remarks	Text
Language	Text
Orientation	Integer (from -2 to +2)
Evaluation	Integer (from -2 to +2)
Confidence	Integer (from 1 to 5)

I will need to import the JSON and run some data cleaning on it to determine the number of NA values and other data preparation tasks specific to this dataset. However, I will certainly need to remove stopwords in two languages. Symbols, like emoticons, are not likely to be a factor.

The columns most likely to be included in the sentiment analysis are the Language and Evaluation columns, where Language is the input text and Evaluation is the label. I will also review the Text and Remarks columns for additional text and the Orientation and Confidence columns for additional classification.

Functionalities

Some of the key NLP techniques this analysis will employ are:

1. Sentiment Analysis – This is the crux of the project.
2. Lemmatization and stemming – To properly prepare the data for comparison
3. Keyword Extraction – To identify representative keywords that could determine overall sentiment.
4. Named Entity Recognition – To filter out proper names and make the sentiment analysis more objective.

I will attempt to include the following user interaction features:

1. Users will be able to filter by positive or negative sentiments.

2. Users will also be able to search for a specific author or reviewer and specific keywords in either the topic or the review

Communication and Sharing

I have created a public GitHub with a ReadMe and the initial dataset. I will publish the project to this repository as I create it:

https://github.com/perezrd5/NLPAdademicSentimentAnalysis

Personal Contribution Statement

The concept and all of the work involved are my own work without the assistance of others.

References

Paper Reviews Data Set. UCI Machine Learning Repository: Paper Reviews Data Set. (n.d.). Retrieved October 9, 2022, from https://archive.ics.uci.edu/ml/datasets/Paper+Reviews