APPLYING MACHINE LEARNING TO STUDENT FEEDBACK THROUGH SENTIMENT ANALYSIS

Eric Andersson, Christopher Dryden, Chirag Variawa

University of Toronto

eric.andersson@mail.utoronto.ca, chris.dryden@mail.utoronto.ca, chirag.variawa@mail.utoronto.ca

Abstract –Machine learning is used to analyze student feedback in first-year engineering courses. This exploratory work builds on previous research at the University of Toronto, where a multi-year investigation used an online survey to collect quantitative and qualitative data from incoming first-year students. [1] (N ~1000)

Sentiment analysis, a machine learning method, is used to investigate the relationship between hours of study outside of scheduled instructional hours and qualitative survey feedback sentiment. The results are visualized with chronological sentiment graphs, which contextualize the results in relation to key events during the school year.

Large drops in sentiment were seen to occur during weeks with major assessments and deadlines. An inverse correlation between hours spent outside of class and feedback sentiment was also noticed.

Keywords: Sentiment analysis, workload, machine learning, natural language processing, engineering education

1. INTRODUCTION

1.1. Background

Many engineering programs solicit feedback from their students with the goal of improving coursework and the student experience. Feedback is often collected in the form of surveys. The results of these surveys are large sets of both qualitative and quantitative data.

The motivation for this work is the difficulty encountered when manually analyzing the extensive amounts of qualitative information from student feedback surveys. Manual reviewing of feedback is often impacted by selection bias, further impacting significance and accuracy [2]. The volume of collected data makes manual analysis time consuming and expensive. The lack of efficient methods to show qualitative correlations means that several studies have solely relied on quantitative survey questions that limit the students response to a single metric. [3]

1.2. Problem Definition

Machine learning is a field of computer science that applies data and statistical methods to allow programs to

improve task performance without being explicitly programmed. It is used in a variety of applications such as computer vision, data analytics, and text analysis.

Natural-language processing is the practice of using machine learning to process and evaluate qualitative written information. A well-established area in the practice is sentiment analysis, which is the process to evaluate the sentiment of textual data. Sentiment analysis has been used to evaluate individual university courses by comparing qualitative feedback to class ratings [4], and to chronologically compare qualitative feedback to instructor ratings [5]. Open answer responses often have the potential to provide additional information not directly measured in the survey.

Current literature measuring the perceived workload of first year engineering students relies heavily on directly asking student to rate their perceived workload on a Likert response scale. [1] While many studies include qualitative response sections, few studies quantitatively measure that data. The problem definition that this paper focuses on is: *Can machine learning methods such as sentiment analysis capture meaningful information of the first-year undergraduate engineering experience?*

This question provides an opportunity to use sentiment analysis to produce exploratory work. A standard for comparison has been derived from the work done by Darlee Gerrard et al. [1], where student workload was measured chronologically using a survey. This research uses results from a similar survey to investigate a relationship between hours of study outside of scheduled instructional hours and qualitative student feedback sentiment.

1.3. Selected Methodology

Many approaches of using natural-language processing were explored to gain a better understanding of the first-year experience, such as sentence topic identification and clustering. However, many of these methods were ineffective due to the lack of data and the high variability of the topics in the written responses.

The approach of using machine learning to augment the ability to measure perceived workload through qualitative text responses was guided by the requirement of most machine learning models to have very large amounts of labeled data to create meaningful results. This limited the scope to machine learning methods that had pre-existing

models that did not require large amounts of labeled data to use.

To conduct the experiments, Microsoft Azure Machine Learning Studio was used due to the operational simplicity it affords. Microsoft has extensively trained algorithms, which eliminates the need to provide a training dataset. Additionally, the software service utilizes a graphical programming interface, allowing for rapid prototyping and iterating. Azure uses data augmentation techniques such as Term Frequency Inverse-Document Frequency (TFIDF), tokenizing, word rooting, and sentence hashing due to the relatively small amount of data used [6].

2. METHODOLOGY

Figure 1 outlines the parallel processes used to process the data. Qualitative data (i.e. text responses) is processed using the left branch, while quantitative data (i.e. numerical responses) is processed with the right branch.

Figure 1. Flowchart of the processing paths for qualitative and quantitative data from the First Year Survey.

2.1. Weekly First Year Survey

The First Year Survey is run by the First Year Office of the Faculty of Applied Science at the University of Toronto. This multi-year weekly survey is sent to first-year engineering students who are taking six courses in the fall semester, and six in the winter semester. Lectures in the fall semester for the 2017-2018 academic year, referred to as Fall 2017, ran from September 7th, 2017 to December 6th,

2017. Winter semester classes ran from January 4th, 2018 to April 11th, 2018 and is referred to as Winter 2017.

Data is collected from a survey conducted by the first-year office. Emails are sent to first-year students once per week, inviting them to respond to the online survey. The survey prompts the student to enter identifying information such as their name, student number, and discipline of study. The body of the survey is comprised of a repeated set of four questions for each course they are enrolled in, based on their discipline. The first three questions use a Likert scale rating, encompassing workload and course content difficulty, followed by an open-ended question for course-specific content improvements. At the end of the survey, the student may enter comments into an optional open-ended form. [1]

2.2. Data Collection

Of the questions posed to the students in the weekly survey, results from two categories are chosen for analysis: workload hours and comments.

1. *"For [the course] how much focused time did you spend on all assignments for this course outside of class this week? (open ended, decimal answers accepted)"*
2. *"Do you have any other comments that will help us better understand first-year student workload?"*

The first question is asked for each course the students are enrolled in, based on their discipline of study. Their quantitative responses indicate the number of hours spent outside of class on coursework such as completing assignments, studying for evaluations, etc.

The second question is asked to the students once at the end of the survey. It allows students to express their opinions and feelings on their coursework and workload. These responses are the source of qualitative data for the study. A sample of the full body of responses is shown in Appendix A.1.

2.3. Data Preprocessing

The text responses being analyzed are general comments without a specific focus beyond first-year student workload. This requires that the hours spent outside of class must be combined for each student into an aggregate number, representing the total amount of time each student spends outside of class on coursework. Outliers with extremely high hours (Such as hours exceeding the number of hours in a week) are eliminated.

The general nature of the text responses also means it is possible for the student's responses to contain both positive and negative comments. The text responses are preprocessed with a Python script to separate individual responses by sentence. For the Fall 2017 semester, 97 text responses were split into 486 sentences. For Winter 2017, 54 text responses were split into 156 sentences. Appendix A.2 contains a sample of the full dataset of sentences.

2.4. Data Processing

Figure 2 shows the data processing workflow within Microsoft Azure Machine Learning Studio. At the top, it starts with a data set comprising of a column of dates and a column of sentences. The dataset is uncompressed by the “Unpack Zipped Datasets” module, and the data type of the sentence column is converted into strings by the “Edit Metadata” module in preparation for sentiment analysis. Sentiment analysis is performed in the “Execute Python Script” module on each sentence, and scores for each sentence are added to a new column. (Script code detailed in Appendix C.1) The “Add Columns” module combines the dates, sentiment scores, and sentences into a single dataset, which is saved as a CSV by the “Convert to CSV” module.

Figure 2. Data processing using Azure Machine Learning Studio interface. Sentiment is analyzed in the “Execute Python Script” step.

Microsoft Azure’s Sentiment Text Analytics is a machine learning classification algorithm used to generate a sentiment value between 0 and 1. Values closer to 1 indicate positive sentiment, while values closer to 0 indicate negative sentiment. The model is pretrained with an extensive body of text with sentiment associations. It is not necessary to provide training data. The model uses a combination of techniques during text analysis, including text processing, part-of-speech analysis, word placement, and word associations.

Sentiment analysis is performed on the entire sentence, as opposed to extracting sentiment for individual words in the text. In practice, there is a tendency for scoring accuracy to improve when sentences are shorter and have common words. During an objectivity assessment phase, the model determines whether a document is objective or contains sentiment. If no sentiment is detected, then the

response is rated 0.50. Appendix A.3 contains a sample of processed data.

2.5. Data Postprocessing

After processing in Azure, the exported CSV file containing sentiment scores is opened in Excel. The file contains three columns: Sentiment, Date, and Comment. Because sentiment scores of 0.5 indicate objective comments, rows with a score of 0.5 are removed. A pivot table is created, allowing the sentiment scores to be averaged by week.

The data for hours spent outside of class is inserted into a second pivot table and averaged by week.

2.6. Visualization

The pivot tables are combined into a single table, allowing for average hours spent outside of class and average sentiment scores to be graphed by week on a dual axis line chart.

3. RESULTS

3.1. Fall 2017

Data for the semester of Fall 2017 starts on the 15th of September 2017 and ends on the 6th of December. Results are visualized in Figure 3. Appendix B.1 contains a table of the numerical values. The solid black line indicates the average sentiment score of all response sentences for each week, measured using the left vertical axis. The dashed grey line represents the average number of hours spent outside of class for each week, measured with the right vertical axis. The horizontal axis indicates the week of the semester.

Figure 3. Average sentiment and average hours outside of class, Fall 2017 semester

3.2. Winter 2017

Data for the Winter 2017 semester starts on the 12th of January 2018 and ends on the 6th of April. Results are visualized in Figure 4. Appendix B.2 contains a table of the numerical values. The solid black line indicates the average sentiment score of all response sentences for each week, measured using the left vertical axis. The dashed grey line

represents the average number of hours spent outside of class for each week, measured with the right vertical axis. The horizontal axis indicates the week of the semester.

There is a gap in the sentiment scores for the week of February 9th due to a lack of responses that contained sentences with sentiment, represented by the dotted black line.

Figure 4. Average sentiment and average hours outside of class, Winter 2017 semester

4. DISCUSSION

A preliminary indicator of the ability for sentiment to accurately measure the perceived workload was to compare the average number of hours that students had of workload in a given week. All student's responses were run through the sentiment analysis program and the average weekly sentiment was compared chronologically with the average weekly hours of workload measured in the survey.

In both semesters, average sentiment starts slightly above neutral (0.5), then drops within two weeks, which may be explained by the students having come back from a long break from their studies. This is indicated by label 1 in Figure 5 and Figure 6.

Figure 5. Key events during the Fall 2017 semester.

In the Fall 2017 semester, the following key events are seen to influence average sentiment and hours spent outside of class, indicated in Figure 5 by each label:

2. The lowest number of hours spent outside of class occurred during the week of October 27th, which was during the "Assessment Free Week", in which no marks were given.

3. A major drop in average sentiment can be seen between the weeks of November 3rd and 10th. This is likely due to the three midterms scheduled for that week, with a total of 85% to 90% of grades assigned.
4. Low sentiment continues through the next week, where a major group deliverable was due.
5. Sentiment rises towards the end of the semester, possibly as a result of students having completed major deliverables.

Figure 6. Key events during the Winter 2017 semester

In the Winter 2017 semester, the following key events are indicated by the labels in Figure 6:

2. No average sentiment score for the week of February 9 due to lack of non-objective responses.
3. The lowest sentiment for the winter semester was during the week of March 9th, which also had the highest average hours outside of class for the semester. A major deliverable for Engineering Strategies and Practice II, a core first year course, was due that week, which could explain both the rise in workload and drop in sentiment.
4. Average sentiment rises at the end of the semester, dropping slightly in the last week. The drop can be attributed to the sudden rise in workload.

4.1. Limitations

The variability of responses combined with low of amount of responses relative to the total amount of students limited our ability to effectively measure overall student sentiment. This is especially eminent during the week of February 9th, 2018, where there were no text responses. The Winter 2017 semester had overall lower survey participation.

A limitation of the methodology is the inability to discern sentences that have alternate meanings. Sarcasm is an example of this limitation, as a sarcastic sentence with a negative sentiment would register as a low sentiment even if the intended sentiment is positive.

In a similar manner, words that are spelled incorrectly and sentences that have incorrect grammar structure will

have misinterpreted sentiment due to their literal nature of processing sentences. This is especially problematic in the first-year engineering workload surveys as many students read, write, and speak English as a second language.

5. CONCLUSIONS & FUTURE WORK

5.1. Relationship Between Sentiment and Workload

While there was a moderate negative correlation between sentiment and perceived workload and many of its established indicators, the correlation is not substantive enough to be able to accurately measure perceived workload through sentiment.

If a neural network model was trained to measure workload perception instead of sentiment the model would be much more accurate in its ability to assess workload perception. While reviewing the sentiment scores that did not match up there were various examples of phrases that indicated that the workload was high but were written with positive sentiment terms.

5.2. Effect on Future Survey Design and Data Collection Methods

The methodology of this research is minimally labor intensive, and it opens the possibility of having larger magnitude of analyzed written responses in surveys. This process can allow for the simplification of surveys by replacing many questions with a single written response.

An area of future research that is used quite extensively by engineering education researchers is focus groups. The use of voice to text transcription software and sentiment analysis would allow for education researchers to quantitatively measure sentiment in focus groups with minimal effort or bias.

Acknowledgements

We would like to extend our gratitude to The First Year Office, and Darlee Gerrard for laying the groundwork upon which this research was built.

References

- [1] Darlee Gerrard, Kirstin Newfield, Narges Balouchestani Asli, and Chirag Variawa, "Are Students Overworked? Understanding the Workload Expectations and Realities of First-Year Engineering" in *Proc. ASEE Annual Conference & Exposition*, (Columbus, OH; 25-28 June 2017), 21 pp., 2017. Available as of May 4, 2018 at <https://www.asee.org/public/conferences/78/papers/18877/view>
- [2] Andrew Bacher-Hicks, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger, "An evaluation of bias in three

measures of teacher quality: value-added, classroom observations, and student surveys," *NBER Working Papers*, 47 pp., 2017. Available as of May 5, 2018 from

https://scholar.harvard.edu/files/andrewbacherhicks/files/an_evaluation_of_bias_in_three_measures_of_teacher_quality.pdf

- [3] Steward H. Oppong, "The problem of sampling in qualitative research," *Asian Journal of Management Sciences and Education*, vol 2, no. 2, pp. 202-210, 2013. Available as of May 4, 2018 at [http://www.ajmse.leena-luna.co.jp/AJMSEPDFs/Vol.2\(2\)/AJMSE2013\(2_2-21\).pdf](http://www.ajmse.leena-luna.co.jp/AJMSEPDFs/Vol.2(2)/AJMSE2013(2_2-21).pdf)
- [4] Quratulain Rajput, Sajjad Haider, and Sayeed Ghani, "Lexicon-based sentiment analysis of teachers' evaluation," *Applied Computational Intelligence and Soft Computing*, vol. 2016, 12 pp., 2016. Available as of May 4, 2018 at <https://www.hindawi.com/journals/acisc/2016/2385429/cta/>
- [5] Sujata Rani and Parteek Kumar, "A sentiment analysis system to improve teaching and learning," *Computer*, vol. 50, no. 5, pp. 36-43, 2017. Available as of May 4, 2018 at <https://ieeexplore.ieee.org/document/7924253/>
- [6] Nabeela Altrabsheh Mohamed Medhat Gaber, and Mihaela Coccea, "SA-E: Sentiment analysis for education," in *5th KES International Conference on Intelligent Decision Technologies*, (Sesimbra, Portugal; 26-28 June 2013), 10 pp., 2013. Available as of May 4, 2018 at [https://researchportal.port.ac.uk/portal/en/publications/sae-sentiment-analysis-for-education\(c99ed217-4c60-494f-8c2c-4f049cf43585\).html](https://researchportal.port.ac.uk/portal/en/publications/sae-sentiment-analysis-for-education(c99ed217-4c60-494f-8c2c-4f049cf43585).html)

APPENDIX A: DATA

A.1 Raw Data

A sample of the data as exported from the survey results. The sample is from Fall 2017.

DateTime	Q60_1OT
9/15/2017 4:13:18 PM	The way my schedule is, we only have one night to finish
9/15/2017 4:42:39 PM	Students go from summer to frosh week to feeling bomb
9/15/2017 6:18:49 PM	The workload is fairly acceptable for engineering student
9/15/2017 6:38:29 PM	For MAT186 and MAT188, there was too much review of t
9/15/2017 7:39:42 PM	More elaboration on new lessons!
9/15/2017 9:21:06 PM	Need more help with time management skills.
9/16/2017 1:14:12 AM	Making blackboard information more clear and distinct so
9/16/2017 11:30:00 AM	On its own the workload would be challenging but manag
9/16/2017 11:43:32 AM	I am not used to the current schedule on Friday and was i
9/16/2017 1:17:19 PM	Although the first-year student workload is relatively less
9/16/2017 7:16:09 PM	There seems to be an overwhelming amount of work you

A.2 Preprocessed Data

A sample of preprocessed text responses. The sample is from Fall 2017.

Date	Sentence
15/09/2017	The way my schedule is, we only have one night to finish the marked prob
15/09/2017	However, other disciplines have the weekend to complete these problem
15/09/2017	Students go from summer to frosh week to feeling bombarded pretty quik
15/09/2017	Mostly because we don't have set schedules so we don't really know what
15/09/2017	If there was a way to help introduce students to ease into what university
15/09/2017	The workload is fairly acceptable for engineering student.
15/09/2017	For MAT186 and MAT188, there was too much review of the regular Ontari
15/09/2017	More elaboration on new lessons!
15/09/2017	Need more help with time management skills.
16/09/2017	Making blackboard information more clear and distinct so we don't fall be

A.3 Processed Data

A sample of data as exported from Microsoft Azure Machine Learning Studio. The sample is from Fall 2017.

SentimentScore	Date	Text
0.101800144	15/09/2017	The way my schedule is, we only have one nigt
0.171418458	15/09/2017	However, other disciplines have the weekend
0.058541924	15/09/2017	Students go from summer to frosh week to fee
0.055066615	15/09/2017	Mostly because we don't have set schedules sc
0.93201828	15/09/2017	If there was a way to help introduce students t
0.960497379	15/09/2017	The workload is fairly acceptable for engineeri
0.5	15/09/2017	For MAT186 and MAT188, there was too much r
0.83629787	15/09/2017	More elaboration on new lessons!
0.23356995	15/09/2017	Need more help with time management skills.

APPENDIX B: RESULTS

B.1 Fall 2017 Results

Week	Average Sentiment	Average Hours
September 15	0.52	21.3
September 22	0.39	32.1
September 29	0.37	36.7
October 06	0.47	30.8
October 13	0.48	29.2
October 27	0.48	10.0
November 03	0.42	35.6
November 10	0.23	38.2
November 17	0.22	34.2
November 24	0.24	46.2
December 01	0.53	42.2

B.2 Winter 2017 Results

Week	Average Sentiment	Average Hours
January 12	0.50	14.8
January 19	0.56	20.6
January 26	0.23	18.2
February 2	0.43	22.3
February 9	N/A	15.0
February 16	0.45	28.5
February 23	0.33	29.2
March 2	0.35	22.8
March 9	0.10	45.5
March 16	0.33	29.6
March 30	0.45	20.1
April 6	0.35	40.5

APPENDIX C: CODE

C.1 Data Processing Script

```
import urllib2
import urllib
import sys
import base64
import json
import numpy as np
import pandas as pd

# The entry point function can contain up to two input arguments:
# Param<dataframe1>: a pandas.DataFrame
# Param<dataframe2>: a pandas.DataFrame
def azureml_main(dataframe1 = None, dataframe2 = None):
    # Execution logic goes here
    account_key = str(dataframe2['Col1'][0])
    base_url = 'https://westus.api.cognitive.microsoft.com/'
    headers = {'Content-Type': 'application/json', 'Ocp-Apim-Subscription-Key': account_key}
    sentiment_scores = []
    num_examples = len(dataframe1.index)
    input_texts = '{"documents":['
    #for each record
    for i in range(0, num_examples):
        input_text = str(dataframe1['Text'][i])
        input_text = input_text.replace("\n", "")
        input_texts = input_texts + '{"id":"' + str(i) + '", "text":"' +
        input_text + '"}, '
    input_texts = input_texts + ']'
    print input_texts
    # Detect sentiment.
    batch_sentiment_url = base_url + 'text/analytics/v2.0/sentiment'
    req = urllib2.Request(batch_sentiment_url, input_texts, headers)
    response = urllib2.urlopen(req)
    result = response.read()
    obj = json.loads(result)
    for sentiment_analysis in obj['documents']:
        sentiment_scores.append( str(sentiment_analysis['score']))
    sentiment_scores = pd.Series(np.array(sentiment_scores))
    df1 = pd.DataFrame({'SentimentScore': sentiment_scores})
    # Return value must be of a sequence of pandas.DataFrame
    return df1
```