Sentiment Analysis of Student Surveys - A Case Study on Assessing the Impact of the COVID-19 Pandemic on Higher Education Teaching *

Haydée G. Jiménez,
Marco A. Casanova
Departamento de Informática
PUC-Rio
Rio de Janeiro, Brazil
{hjimenez,
casanova}@inf.puc-rio.br

Anna Carolina Finamore
COPELABS
Lusófona University
Lisboa, Portugal
anna.couto@ulusofona.pt

Gonçalo Simões
Google Research
United Kingdom
gsimoes@google.com

ABSTRACT

Sentiment Analysis is a field of Natural Language Processing which aims at classifying the author's sentiment in text. This paper first describes a sentiment analysis model for students' comments about professor performance. The model achieved impressive results for comments collected from student surveys conducted at a private university in 2019/20. Then, it applies the model to different scenarios: (i) in-person classes taught in 2019 (pre-COVID); (ii) the emergency shift to online, synchronous classes taught in the first semester of 2020 (early-COVID); and (iii) the planned online classes taught in the second semester of 2020 (late-COVID). The results show that students acknowledged the effort professors did to keep classes running during the first semester of 2020, and that the enthusiasm continued throughout the second semester. Furthermore, the results show that students evaluated professors' performance for online courses better than for in-person courses.

Keywords

sentiment analysis, BERT, online classes, in-person classes

1. INTRODUCTION

The systematic evaluation of a Higher Education Institution (HEI) provides its administration with valuable feedback about several aspects of academic life, such as the reputation of the institution and the individual performance of faculty. In fact, in some countries, it is mandatory that HEIs implement self-evaluation committees, whose members are elected by the various segments of the community and whose duties include the preparation of annual reports assessing the performance of the institution on predefined aspects.

In particular, student surveys are a first-hand source of information that help assess professor performance and course adequacy. Such surveys are typically organized as a questionnaire with *closed-ended* questions, which the student answers by choosing predefined alternatives, and *open-ended* questions, which the student answers by freely writing comments on the topic of the question. Albeit interesting and useful, the analysis of open-ended questions poses challenges, such as how to summarize the comments and how to determine the sentiment of the comments.

The primary goal of this paper is to introduce a sentiment analysis model for students' comments in the context of questionnaires designed to assess professor performance, and to evaluate the model using data from student surveys applied at Brazilian University in 2019 and 2020.

Studying this particular period of time is interesting because, in early 2020, the COVID-19 pandemic forced the Brazilian University to move all classes online, taught with the help of a videoconferencing software and a Learning Management System (LMS), and they so remained throughout 2020. This change in instructional model offers the unique opportunity to compare the in-person classes in 2019 (pre-COVID scenario), with the emergency shift to online, synchronous classes in the first semester of 2020 (early COVID scenario), and with the planned online classes in the second semester of 2020 (late-COVID scenario). Therefore, the second goal of this paper is to apply the sentiment analysis model developed to the case study data to compare the overall sentiment of the students' comments about professor performance in these different scenarios.

The results reported in this paper indicate that the sentiment analysis model developed achieves good performance in the classification of the sentiment expressed by the students' comments about professor performance. This model was separately applied to the different scenarios covered by the case study data. The results show that students acknowledged the effort professors did to keep classes running during the first semester of 2020 (early-COVID sce-

*(Does NOT produce the permission block, copyright information nor page numbering). For use with ACM_PROC_ARTICLE-SP.CLS. Supported by ACM.

Haydée Guillot Jiménez, Anna Carolina Finamore, Marco Antonio Casanova and Gonçalo Simões "Sentiment Analysis of Student Surveys - A Case Study on Assessing the Impact of the COVID-19 Pandemic on Higher Education Teaching". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 353-359. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

nario), and that the enthusiasm continued throughout the second semester of 2020 (late-COVID scenario). These conclusions are justified by the peak in positive comments observed in the first semester of 2020, as compared with the other semesters. Furthermore, the results show that students evaluated professor performance for online classes better than for in-person classes. To a large extent, these remarks are consistent, for example, with the findings of a random-sample survey, conducted in late May 2020, involving more than 1,000 US college students whose classes moved from in-person to completely online in early 2020 [13]. However, they have to be cross-checked with other surveys conducted in 2019/2020 at the Brazilian University and elsewhere.

The remainder of this paper is organized as follows. Section 2 summarizes related work. Section 3 presents the case study used in the paper. Section 4 details the model for sentiment analysis. Section 5 describes the results obtained with the case study data. Section 6 contains the conclusions and directions for future work.

2. RELATED WORK

Sentiment Analysis (SA), also known as Opinion Mining, is a field of natural language processing (NLP) where the main focus is to automatically analyze people's opinions and sentiments [11]. According to Pang and Lee [15], for most of us, the decision-making process takes into consideration "what other people think". Based on this assertion, it is easy to understand why SA is very popular in several domains, such as tourism, restaurants, movies, music, and, more recently, education.

Chaturvedi et al. [5] addressed the essential task of eliminating "real" or "neutral" comments that do not express a sentiment. The article reviewed hand-crafted and automatic models for detecting subjectivity in the literature, comparing the advantages and limitations of each approach. Ahuja et al. [1] addressed the analysis of comments from one of the most popular Twitter platforms. As the comments are not structured, they used six techniques to pre-process the comments. They then applied two techniques (TF-IDF and N-Grams) to classify comments, and concluded that the TF-IDF word level of sentiment analysis is 3-4% higher than the use of N-characteristics. Prusa et al. [17] also concentrated on Twitter data. They analysed the impact of ten filter-based feature selection techniques on the performance of four classifiers. Nazare et al. [14] analyzed about 1,000 Twitter comments using various machine learning approaches, separately or in combination, to classify the comments. Unlike other articles with traditional approaches to analyze the sentiment of short texts, Li and Qiu [10] did not consider the relationship between emotion words and modifiers, but they showed how to mitigate these problems through the sentiment structure and rules that captured the text sentiment. The results of an experiment with microblogs validated the efficacy of their approach.

Analyzing comments from sales Web sites is important to detect if users are praising or criticizing the products they consume. Bansal and Srivastava [4] used the word2vec model to convert comments into vector representations using CBOW (continuous bag of words), which were fed to a classifier. Experimental results showed that Random Forests using CBOW achieved the highest precision. Khoo and Johnkhan [9] analysed comments from the Amazon Web site, using a new general-purpose sentiment lexicon, called WKWSCI Sentiment Lexicon, and compared it with five existing lexicons. Akhtar et al. [2] used classification algorithms, like Conditional Random Filed (CRF) and Support Vector Machine (SVM), to classify comments from different Indian Web sites.

Zhou and Ye [22] reviewed journal publications between 2010-2020 in SA applied to the education domain and, among others future research directions, they pointed out: (i) the need to explore SA in the learning cross-domain; (ii) consider a combination of text mining and qualitative answers (questionnaires or interviews) to understand the psychological motivation behind learning sentiment; (iii) explore the association between sentiment, motivation, cognition, and also demographic characteristics to regulate the emotions of learners. Santos et al. [19] studied SA in online students' reviews to identify factors that influence international students' choice for a HEI. They also suggested aspects that HEI managers may have to consider to attract more international students, such as: online information about (HEI) offerings, students' comments about their experiences, international environment, courses taught in English, and support to students' accommodation or expenses. Sindhu et al. [20] proposed an aspect-oriented SA system based on Long Short-Term Memory (LSTM) models. They considered two datasets with students' comments, namely: the Sukkur IBA University and a standard SemEval-2014. They suggested that the evaluation of teaching performance would have to consider six dimensions: teaching pedagogy, behavior, knowledge, assessment, experience, and general. We previously created a tool for the analysis of student comments [8] but it was limited to a fixed, manually created dictionary, which might therefore not take into account some relevant words.

The choice of a university to enroll in is a difficult decision and, at the same time, the information available on the internet is overwhelming. To address these issues, Balachandran and Kirupananda [3] proposed an aspect-based sentiment analysis tool to evaluate the reputation of universities in Sri Lanka from users' comments in Facebook and Twitter, using the StanfordCoreNLP library to perform sentiment analysis. Lytras et al. [12] built the Learning Analytics Dashboard for E-Learning (LADEL) tool to monitor different sources, such as student blogs, social networks and Massive Open Online Courses (MOOC) in search of comments that express satisfaction, anxiety, efficiency, frustration, abandonment. LADEL is composed of four modules: collection, cleaning, word cloud and sentiment of opinion. Sivakumar and Reddy [21] extracted students' comments using the Twitter API and tried to analyze the relations between word aspects and phrases of student opinion. They used a sentiment package available in R to find the polarity of the sentences and then applied k-mean clustering and naïve Bayes for the sentiment analysis classification.

de Oliveira and de Campos Merschmann [6] analyzed the combination of NLP pre-processing tasks (tokenization, POS tagging, stemming, among others) with three classifiers (Ran-

dom Forest, Support Vector Machine, and Multilayer Perceptron), and discussed their predictive performance. They evaluated these tasks in five Portuguese datasets related to sentiment analysis, encompassing comments, news and tweets. They analyzed some combinations of preprocessing tasks and classifier

This paper focuses on identifying students' sentiments expressed in comments about professor performance in Higher Education. It uses the pre-trained model called Bidirectional Encoder Representations from Transformers (BERT) [7] for the sentiment analysis task. BERT-style models are the current state-of-the-art in several NLP tasks, including entity recognition and sentiment analysis. BERT's architecture is based on multi-layered transformers, which are particularly optimized to be trained on GPUs and TPUs with significant amounts of data. For this reason, a recipe for success with these models is to pre-train them with large datasets (in the order of millions of documents) on general tasks such as masked language models or next sentence predictions [7]. This pre-training allows the model to learn a lot about some language patterns (that are independent of the task we care about) and make it easier to train them specifically for other language tasks even without the need for large amounts of annotated data. Our corresponding code is available at GitHub[1].

3. CASE STUDY

3.1 Course Survey Data

In the rest of this paper, we use *course* to denote "a series of lectures in a particular subject", and *class* to describe "a particular instance of a course". Therefore, students enroll in a class of a course. We assume that classes run on a per semester basis, and use <year>.1 and <year>.2 to denote the first and second semesters of the calendar year, respectively.

Since 2005, at the Brazilian University used in the case study, students are invited, at the end of each semester, to answer a questionnaire for each class they took in the semester. Students' participation in the survey is not mandatory. The questionnaire has a set of closed-ended questions about the professor that taught the class, and a separate set of closed-ended questions about the course the class is an instance of. For each closed-ended question, the student chooses a score from a Likert scale (1-5). The questionnaire also has one open-ended question which invites students to write as many sentences as they like to express their evaluation of the professor that conducted the class, and likewise for the course the class is an instance of. The comments are in Portuguese and the sentences are often ungrammatical. We are interested in the sentiment analysis of the students' comments about the professor performance, which we will refer to as the *comments* for brevity.

The purpose of the case study is to analyse comments collected from the questionnaires applied in the first and second semesters of 2019 and 2020. However, we also use the comments collected from the questionnaires applied in both semesters of 2018 for pre-training (see Section 5). The reason for using comments from 2018 for pre-training is that we wanted to make sure that no comment used in the analysis step has been observed before in the pre-training step. Using the 2018 data is possible because it has been observed that the vocabulary students use to write comments has not changed significantly over the years. Table 1 presents the number of comments for the 2018, 2019 and 2020 student surveys.

[1]https://github.com/hguillot/Sentiment-Analysis-of-Student-Surveys-with-BERT

Table 1: Number of comments about professor performance in classes.

Semester	#Comments
2018.1 and 2018.2	10,077
2019.1	3,182
2019.2	1,910
2020.1	3,492
2020.2	2,219

Table 2: Structure of the professor questionnaires.

Year	Class Mode	#Closed-ended Questions	#Open-ended Questions
2018	in-person	10	1
2019	in-person	16	1
2020	online	20	1

As far as professor evaluation is concerned, the questionnaires varied slightly from 2018 to 2019. Also, in early 2020, the COVID pandemic forced the university to move all classes online, taught with the help of a videoconferencing software and a Learning Management System (LMS), and they so remained throughout 2020. The questionnaire, used for classes taught in 2020.1 and 2020.2, was then modified accordingly. Table 2 summarizes the structure of the various questionnaires that the case study is concerned with.

Given this new reality, forced by the COVID pandemic, it is reasonable to ask if the professors were prepared for online classes and if this would affect the students' evaluation of the professor performance at the end of 2020.1 (the early-COVID scenario).

As a simple answer to this conjecture, consider the last closed-ended question incorporated in the 2019/20 surveys: *"O: Overall evaluation of the professor"*. Figure 1 depicts the distribution of the scores of Question O per semester, grouped as 1 and 2, for "negative", 3, for "neutral", and 4 and 5, for "positive", considering only questionnaires with a non-empty comment about professor performance. Figure 1 shows that students in fact evaluated the overall professor performance better in 2020.1 (again, the early-COVID scenario) than in the other semesters.

But the question remains if the overall sentiment of the comments about professor performance points in the same direction.

3.2 Use of the Course Survey Data

This section describes how the course survey data were used to construct models for the sentiment analysis of comments about professors performance (recall that each questionnaire

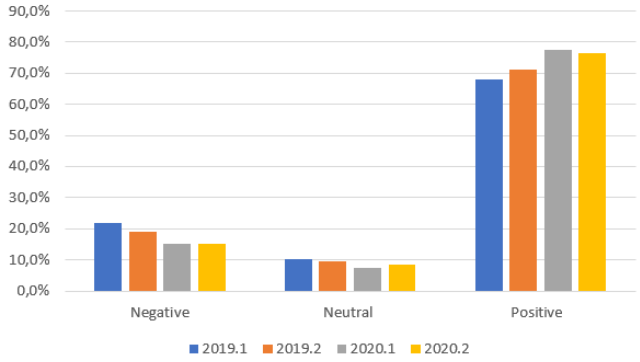

Figure 1: Distribution of the scores of Question O per semester (considering only questionnaires with a non-empty comment about professor performance).

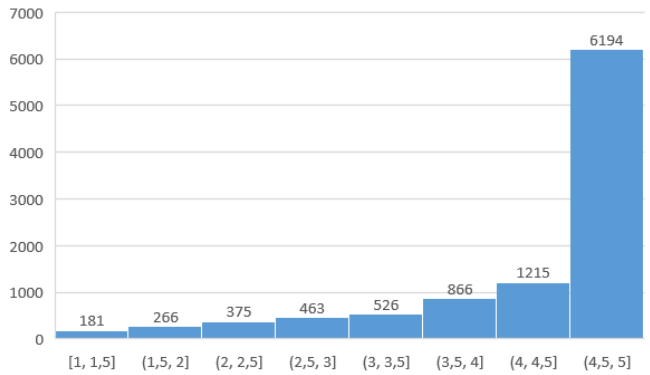

Figure 2: Distribution of the average score of all questions of a questionnaire from 2018.

has only one such comment).

We first observe that no text pre-processing was necessary, as in the Twitter sentiment analysis reported in [16], since the students' comments do not significantly depart from written Portuguese, albeit they often contain ungrammatical sentences.

The models used manually annotated comments, obtained as follows. From the course surveys of the two semesters of 2019, 800 questionnaires with non-empty professor comments were randomly chosen, using the following criteria: 5 samples were chosen for each of the Likert scale scores (1-5) for each of the 16 closed-ended questions (5 * 5 * 16 = 400) in each of the semesters (400 * 2 = 800). The comments of the selected questionnaires were manually classified into 3 categories: *positive*, when the comment only praised the professor; *negative*, when the comment only criticized the professor; and *neutral* when the comment expressed no opinion or when the comment both praised and criticized the professor. Table 3 shows the number of comments in each of these classes.

The pre-training step (see Section 5) used data from the 2018 student surveys as follows. We considered a dataset with all questionnaires with non-empty comments from the 2018 student surveys. But, since the questionnaire applied in 2018 had no overall professor evaluation (Question O), we used the average score $s_{avg}[q] \in [1, 5]$ of all questions of a questionnaire q to induce a label $c[q] \in \{$"*negative*", "*neutral*", "*positive*"$\}$ for the comment as follows: if $s_{avg}[q] < 3$ then $c[q] =$ "*negative*"; if $3 \leq s_{avg}[q] < 4$ then $c[q] =$ "*neutral*"; and if $s_{avg}[q] \geq 4$ then $c[q] =$ "*positive*". Figure 2 shows the distribution of the average scores obtained.

4. A SENTIMENT ANALYSIS MODEL

In the paper, we focus on the *polarity classification* task, whose focus is to classify comments, which express opinions or reviews, into "positive", "negative" or "neutral", or even into more than these three classes. We neither consider *subjectivity classification*, i.e., the task of verifying the subjectivity and objectivity of a comment, nor *irony detection*, i.e., the task of verifying whether the comment is ironic or not.

Table 3: Distribution of the number of questionnaires per class of comment about professor performance, using the manual classification and the automatic classification induced by the score of Question O (considering 800 questionnaires with a manually classified comment about professor performance).

Year	Classification	Positive	Negative	Neutral
2019.1	Manual	107	220	73
	Automatic	187	150	63
2019.2	Manual	119	203	78
	Automatic	201	138	61

We use BERT [7], which achieves outstanding results on a number of NLP tasks. The core of the architecture has been pre-trained on a very large amount of unlabeled data. The model is then fine-tuned on small supervised datasets, designed for the task in question.

For our case study, BERT encodes each comment into a 768-dimensional embedding and, then, a dense layer transforms the embeddings into a three-dimensional vector for each comment that indicates the probability that the comment belongs to each of the three classes - "positive", "negative" or "neutral". We adopted the BERT-Base, Multilingual Cased version[2] (for 104 languages, with 12-layer, 768-hidden, 12-heads, 110M parameters), which is required since the comments are written in Portuguese. In order to significantly speed up the training and inference with our model, we limited the size of each input comment to 64 tokens, which is enough to cover the vast majority of the comments. Any comment with less than 64 tokens was padded with the '[PAD]' symbol already allocated in BERT's vocabulary and any comment with more than 64 tokens was truncated.

Finally, the model was implemented using KERAS and running on GPU's.

5. EXPERIMENTS AND RESULTS

We started our experimental setup by executing a pre-training step that aims at getting the model used to the style of stu-

[2]Available at `https://github.com/google-research/bert/blob/master/multilingual.md`

Table 4: Results of the experiments.

Experiment	Accuracy	Precision	Recall	F1
Zero-shot	50.2±2.3	54.2±2.2	51.8±2.8	53.0±2.4
From scratch	86.3±1.8	**84.5±2.3**	83.0±3.1	83.7±2.4
Fine-tuned	**87.5±2.0**	84.6±2.0	**84.8±2.0**	**84.6±2.5**

dent's comments through non-annotated data. In order to do that, we used the set of comments from the 2018 student surveys, and the scores they assigned to the professor as a proxy to the labels, as explained in Section 3.2. We started the pre-training experiment with the multilingual BERT checkpoint that is publicly available and trained for 10 epochs, resulting in a newly trained checkpoint which we call from this point on the *pre-trained checkpoint*.

After the pre-training step, we proceeded to experiment with three setups, using a 5-fold cross-validation strategy, applied to the set of 800 manually classified comments. Therefore, each round of cross-validation used 640 comments for training and 160 comments for testing. The three setups we used were as follows:

- *Zero-shot*: this experiment does not perform any training with the manually classified comments. Instead, it performs inference directly using the pre-trained checkpoint that resulted from the pre-training step on the test set. If this model's performance was good, then it would show that manually annotating comments would not be necessary.
- *From scratch*: this experiment does not use the pre-trained checkpoint that resulted from the pre-training step. Instead, it starts with the multilingual BERT checkpoint and uses the manually classified comments to train and evaluate the model. The objective of this experiment is to understand if the pre-training step is necessary to obtain top-quality results.
- *Fine-tuned*: this experiment uses the pre-trained checkpoint that resulted from the pre-training step and then uses it as the starting point when training with the manually classified documents. This experiment aims at evaluating if combining pre-training and manually annotated comments helps in obtaining top-quality results.

Table 4 shows the results of the 5-fold cross-validation (each cell indicates the average and the standard deviation over the 5 rounds). Observe that the fine-tuned model obtained the best results, which indicates that combining pre-training and manually annotated comments helps in obtaining top-quality results.

We have also computed the Fisher-Irwin test [18], to examine the hypothesis that Fine-tuned model does not have an equivalent classification performance when compared to both Zero-shot and From scratch. For this purpose, we computed the Fisher-Irwin test twice. In the first test, our null hypothesis (*Fine-tuned classifier has a proportion of correct classifications equivalent to the proportion of correct classifications from Zero-shot classifier*) was tested against the alternative hypothesis (*Fine-tuned classifier has a proportion of correct classifications superior to the proportion of correct classifications from the Zero-shot classifier*), and the null hypothesis was rejected for the usual levels of statistical significance (5% and 10%). The same happened in our second test where our null hypothesis (*Fine-tuned classifier has a proportion of correct classifications equivalent to the proportion of correct classifications from From scratch classifier*) was tested against the alternative hypothesis (*Fine-tuned classifier has a proportion of correct classifications superior to the proportion of correct classifications from the From scratch classifier*). Based on this, we can conclude that our results are statistically significant, since our null hypotheses were both rejected for the usual levels of statistical significance (5% and 10%), leading us to accept alternative hypotheses.

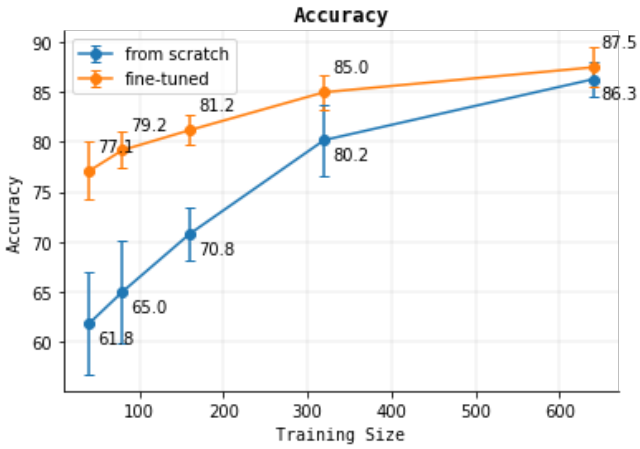

Figure 3: Accuracy for *From scratch* and *Fine-tuned* using train set of 40, 80, 160, 320 and 640 comments.

An important question that arises is about the number of comments that must be manually annotated to achieve an acceptable level of accuracy. To address this question, we ran the following cross validation experiment, with a decreasing number of manually annotated comments used for training. We divided the 800 manually annotated comments into 5 sets of 160 comments each. Let $G_1, ..., G_5$ denote these sets and $\overline{G_i}$ denote the 640 comments not in G_i. For each $i = 1, ..., 5$, we computed the accuracy and the F1-score of the from-scratch and the fine-tuned models, using G_i for testing and subsets of $\overline{G_i}$, of sizes 640, 320, 160, 80, and 40, for training. Finally, for each cardinality of the training sets, we computed the average accuracy and the average F1-score of each model. Figures 3 and 4 depict the results.

Figure 3 shows that, using 640 manually annotated comments for training, the fine-tuned model achieved an average accuracy of 87.5% and the from-scratch model achieved 86.3%, and so on for the other training set cardinalities (320, 160, 80 and 40). Therefore, based on the level of accepted accuracy, one can balance the effort to manually annotate the comments.

Figure 3 also shows that: (i) using just 40 manually annotated comments for training, the fine-tuned model achieved an average accuracy of 77.1%, while the from-scratch model

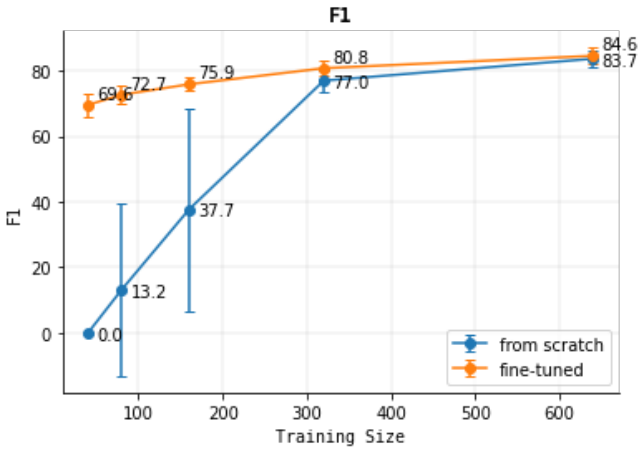

Figure 4: F1 for *From scratch* and *Fine-tuned* using train set of 40, 80, 160, 320 and 640 comments.

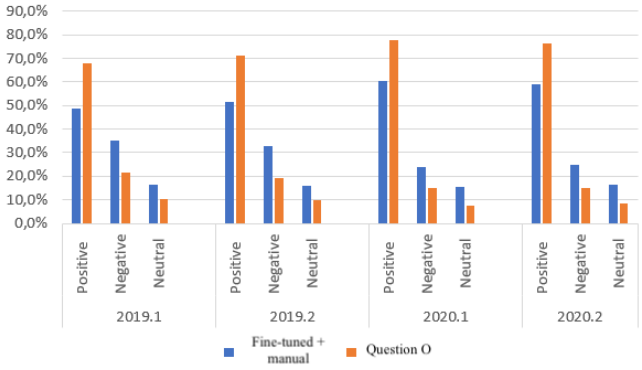

Figure 5: Distribution of the final classification of the comments from all surveys, using the fine-tuned model, added to the manually classified comments from 2019.1 and 2019.2 (shown in blue), and the classification of the comments from all surveys, using the score of Question O (shown in orange).

only achieved an accuracy of 70.8%, when trained with 160 comments, that is, 4 times as much comments; (ii) the fine-tuned model, again trained with just 40 comments, achieved a much better accuracy than that of the zero-shot model, shown in the first line of Table 4 (the zero-shot model is the equivalent to training the fine-tuned model with 0 comments); (iii) the pre-trained check-point had a positive impact, since the fine-tuned curve is always above the from-scratch curve; (iv) the fine-tuned model achieved a standard deviation smaller than that of the from-scratch model, which means that this technique is more stable and less susceptible to changes due to the samples. These observations reinforce that, with an adequate pre-training strategy, we may achieve good results without the need to manually annotate a large amount of data.

Finally, we used the fine-tuned model to classify the full set of comments from the 2020.1 and 2020.2 surveys, and the set of comments from 2019.1 and 2019.2 that were not manually classified. Then, we added the manually classified comments from 2019.1 and 2019.2 to obtain the final distributions for the four semesters, as shown in Figure 5.

For comparison purposes, Figure 5 includes the distributions of the comment classifications induced by the score of Question O as explained in Section 3.1. Note that Question O induces a classification biased towards positive comments, when compared with the classification based on the fine-tuned model. This is also observed when just the manually classified comments are considered.

In conclusion, the distributions of the students' comments sentiment and of the scores of Question O indicate that students evaluated the professor performance better in 2020.1 (the early-COVID scenario) than in the other semesters, which seems to indicate that students acknowledged the effort professors did to keep classes running during 2020.1, and that the enthusiasm continued throughout 2020.2 (late-COVID scenario). Furthermore, students evaluated the professor performance better in 2020.1 and 2020.2 (online classes), by a margin of nearly 10%, when compared with 2019.1 and 2019.2 (in-person classes), respectively.

6. CONCLUSIONS

This paper first described a sentiment analysis model for students' comments about professor performance. The model is based on BERT and has achieved good results when applied to a case study with students' comments about professor performance, obtained in 2019/20.

Then, the paper applied the model to compare the overall sentiment of the students' comments about professor performance in different scenarios: in-person classes in 2019.1 and 2019.2 (pre-COVID scenarios); the emergency shift to online, synchronous classes in 2020.1 (early COVID scenario); and the planned online classes in 2020.2 (late-COVID scenario). The results show that students acknowledged the effort professors did to keep classes running during 2020.1, and that the enthusiasm continued throughout 2020.2. Furthermore, the results show that students evaluated professor performance for online courses better than for in-person courses, by a margin of nearly 10%, which seems to indicate that students favor online classes.

This paper also discussed the number of comments that must be manually annotated to achieve good results. Future experiments can take advantage of this discussion to reduce the manual annotation effort, even with datasets obtained from other universities.

The stability of the models was also investigated, indicating that the fine-tuned model achieved a lower standard deviation, which means that this technique leads to more stable results. The fine-tuned model also achieved a higher performance, when compared to both the zero-shot and from-scratch models, in terms of the proportion of correct classifications, and the difference was statistically significant.

We plan to extend the analysis to past student surveys, which go back to 2005, and to the student survey to be applied at the end of 2021.1, when classes will still be online. We also plan to cross-check these preliminary findings with other surveys conducted in 2019/20 at the Brazilian University and elsewhere.

7. ACKNOWLEDGMENTS

This work was partly funded by FAPERJ under grant E-26/202.818/2017, by CAPES under grant 88882.164913/2010-01, and by CNPq under grants 302303/2017-0 and 162959/2017-6.

8. REFERENCES

[1] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja. The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, 152:341–348, 2019.

[2] M. S. Akhtar, A. Ekbal, and P. Bhattacharyya. Aspect based sentiment analysis in Hindi: Resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2703–2709. European Language Resources Association (ELRA), May 2016.

[3] L. Balachandran and A. Kirupananda. Online reviews evaluation system for higher education institution: An aspect based sentiment analysis tool. *2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pages 1–7, 2017.

[4] B. Bansal and S. Srivastava. Sentiment classification of online consumer reviews using word vector representations. *Procedia Computer Science*, 132:1147–1153, 2018. International Conference on Computational Intelligence and Data Science.

[5] I. Chaturvedia, E. Cambria, R. E. Welsch, and F. Herrera. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77, December 2018.

[6] D. N. de Oliveira and L. H. de Campos Merschmann. Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in brazilian portuguese language. *Multimedia Tools and Applications*, pages 1–22, 2021.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] H. G. Jiménez, M. A. Casanova, B. P. Nunes, and A. C. Finamore. Courseobservatory: Sentiment analysis of comments in course surveys. In M. Chang, D. G. Sampson, R. Huang, A. S. Gomes, N. Chen, I. I. Bittencourt, Kinshuk, D. Dermeval, and I. M. Bittencourt, editors, *19th IEEE International Conference on Advanced Learning Technologies, ICALT 2019, Maceió, Brazil, July 15-18, 2019*, pages 176–178. IEEE, 2019.

[9] C. S. Khoo and S. B. Johnkhan. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511, 2018.

[10] J. Li and L. Qiu. A sentiment analysis method of short texts in microblog. In *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, volume 1, pages 776–779, 2017.

[11] B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[12] M. D. Lytras, E. D'Avanzo, P. Adinolfi, I. Novo-Corti, and J. Picatoste. Sentiment analysis to evaluate teaching performance. *Int. J. Knowl. Soc. Res.*, 7(4):86–107, October 2016.

[13] B. Means and J. Neisler. Suddenly online: A national survey of undergraduates during the covid-19 pandemic, 2020.

[14] S. P. Nazare, P. S. Nar, A. S. Phate, and P. D. D. R. Ingle. Sentiment analysis in twitter. *International Research Journal of Engineering and Technology (IRJET)*, 5(1):880–886, January 2018.

[15] B. Pang and L. Lee. Opinion mining and sentiment analysis (foundations and trends (r) in information retrieval), 2008.

[16] M. Pota, M. Ventura, R. Catelli, and M. Esposito. An effective bert-based pipeline for twitter sentiment analysis: A case study in italian. *Sensors*, 21(1), 2021.

[17] J. D. Prusa, T. Khoshgoftaar, and D. Dittman. Impact of feature selection techniques for tweet sentiment classification. In *FLAIRS Conference*, 2015.

[18] S. M. Ross. *Introduction to probability and statistics for engineers and scientists*. Academic Press, 2020.

[19] C. L. Santos, P. Rita, and J. Guerreiro. Improving international attractiveness of higher education institutions based on text mining and sentiment analysis. *International Journal of Educational Management*, 2018.

[20] I. Sindhu, S. M. Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi. Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation. *IEEE Access*, 7:108729–108741, 2019.

[21] M. Sivakumar and U. S. Reddy. Aspect based sentiment analysis of students opinion using machine learning techniques. In *2017 International Conference on Inventive Computing and Informatics (ICICI)*, pages 726–731. IEEE, 2017.

[22] J. Zhou and J.-m. Ye. Sentiment analysis in education research: a review of journal publications. *Interactive Learning Environments*, pages 1–13, 2020.