Pontifícia Universidade Católica
do Rio de Janeiro

Haydée Guillot Jiménez

On the Processing of Course Survey Comments in Higher Education Institutions

Tese de Doutorado

Thesis presented to the Programa de Pós-graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática.

Advisor : Prof. Marco Antonio Casanova
Co-advisor: Profa. Anna Carolina Finamore do Couto

Rio de Janeiro
November 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Haydée Guillot Jiménez

On the Processing of Course Survey Comments in Higher Education Institutions

Thesis presented to the Programa de Pós-graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática. Approved by the Examination Committee:

Prof. Marco Antonio Casanova
Advisor
Departamento de Informática – PUC-Rio

Profa. Anna Carolina Finamore do Couto
Co-advisor
Universidade Lusófona de Humanidades e Tecnologias

Profa. Carla Faria Leitão
Departamento de Psicologia – PUC-Rio

Profa. Melissa Lemos Cavaliere
TecGraf – PUC-Rio

Profa. Ana Moura Santos
Instituto Superior Técnico – Universidade de Lisboa

Prof. Geraldo Bonorino Xexéo
Universidade Federal do Rio de Janeiro

Rio de Janeiro, November 5th, 2021

Haydée Guillot Jiménez

Graduated in Computer Science from the University of Havana (UH), Havana - Cuba, in 2012. Obtained a Master in Science in Computer Science at the Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro - Brazil, in 2017.

Bibliographic data

Guillot Jiménez, Haydée

On the Processing of Course Survey Comments in Higher Education Institutions / Haydée Guillot Jiménez; advisor: Marco Antonio Casanova; co-advisor: Anna Carolina Finamore do Couto. – 2021.

100 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2021.

Inclui bibliografia

1. Análise de sentimento. 2. Mineração de Dados Educacionais. 3. Visualização de dados. 4. BERT. 5. Resumo de comentários. 6. Similaridade. 7. Entailment. 8. TextRank. I. Casanova, Marco Antonio. II. Finnamore do Couto, Anna Carolina. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

To my loved ones.

Acknowledgments

First of all, I would like to thank my parents for all the love and strength.

To my husband Daniel for supporting me on this trip that ended up being a wonderful experience.

To my brother for always being an example, to my family and true friends in general for helping me to achieve this goal.

Thank you so much to my advisors Marco Antonio Casanova and Anna Carolina Finamore, for all the guidance provided throughout this time of learning.

Then I wish to thank Gonçalo Simoes for bringing BERT to this investigation.

To all my classmates, teachers and staff from the Department of Informatics.

To PUC-Rio and CNPq for funding my research. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Thank you so much to all of you!

Abstract

Guillot Jiménez, Haydée; Casanova, Marco Antonio (Advisor); Finnamore do Couto, Anna Carolina (Co-Advisor). **On the Processing of Course Survey Comments in Higher Education Institutions**. Rio de Janeiro, 2021. 100p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

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 teachers. In particular, student surveys are a first-hand source of information that help assess teacher performance and course adequacy. The primary goals of this thesis are to create and evaluate sentiment analysis models of students' comments, and strategies to summarize students' comments. The thesis first describes two approaches to classify the polarity of students' comments, that is, whether they are positive, negative, or neutral. The first approach depends on a manually created dictionary that lists terms that represent the sentiment to be detected in the students' comments. The second approach adopts a language representation model, which does not depend on a manually created dictionary, but requires some manually annotated test set. The results indicated that the first approach outperformed a baseline tool, and that the second approach achieved very good performance, even when the set of manually annotated comments is small. The thesis then explores several strategies to summarize a set of comments with similar interpretations. The challenge lies in summarizing a set of small sentences, written by different people, which may convey repeated ideas. As strategies, the thesis tested Market Basket Analysis, Topic Models, Text Similarity, TextRank, and Entailment, adopting a human inspection method to evaluate the results obtained, since traditional text summarization metrics proved inadequate. The results suggest that clustering combined with the centroid-based strategy achieves the best results.

Keywords

Sentiment Analysis; Educational Data Mining; Data Visualization; BERT; Comment Summarization; Similarity; Entailment; TextRank.

Resumo

Guillot Jiménez, Haydée; Casanova, Marco Antonio; Finnamore do Couto, Anna Carolina. **Processamento de Comentários de Pesquisas de Cursos em Instituições de Ensino Superior**. Rio de Janeiro, 2021. 100p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A avaliação sistemática de uma Instituição de Ensino Superior (IES) fornece à sua administração um feedback valioso sobre vários aspectos da vida acadêmica, como a reputação da instituição e o desempenho individual do corpo docente. Em particular, as pesquisas com alunos são uma fonte de informação de primeira mão que ajuda a avaliar o desempenho do professor e a adequação do curso. Os objetivos principais desta tese são criar e avaliar modelos de análise de sentimento dos comentários dos alunos e estratégias para resumir os comentários dos alunos. A tese primeiro descreve duas abordagens para classificar a polaridade dos comentários dos alunos, ou seja, se eles são positivos, negativos ou neutros. A primeira abordagem depende de um dicionário criado manualmente que lista os termos que representam o sentimento a ser detectado nos comentários dos alunos. A segunda abordagem adota um modelo de representação de linguagem, que não depende de um dicionário criado manualmente, mas requer algum conjunto de teste anotado manualmente. Os resultados indicaram que a primeira abordagem superou uma ferramenta de linha de base e que a segunda abordagem obteve um desempenho muito bom, mesmo quando o conjunto de comentários anotados manualmente é pequeno. A tese então explora várias estratégias para resumir um conjunto de comentários com interpretações semelhantes. O desafio está em resumir um conjunto de pequenas frases, escritas por pessoas diferentes, que podem transmitir ideias repetidas. Como estratégias, a tese testou Market Basket Analysis, Topic Models, Text Similarity, TextRank e Entailment, adotando um método de inspeção humana para avaliar os resultados obtidos, uma vez que as métricas tradicionais de sumarização de textos se mostraram inadequadas. Os resultados sugerem que o agrupamento combinado com a estratégia baseada em centróide atinge os melhores resultados.

Palavras-chave

Análise de sentimento; Mineração de Dados Educacionais; Visualização de dados; BERT; Resumo de comentários; Similaridade; Entailment; TextRank.

Table of contents

1 Introduction **16**
1.1 Context and Motivation 16
1.2 Problems Addressed and Contributions 16
1.3 Structure 18

2 Related Work **19**
2.1 Sentiment Analysis 19
2.2 Sentiment Analysis in Higher Education 20
2.3 Sentiment Analysis in Portuguese 21
2.4 BERT 22
2.5 Semantics beyond individual sentences 22
2.5.1 Sentence Textual Similarity 22
2.5.2 Recognizing Textual Entailment 23
2.6 Summarization 24

3 Background **25**
3.1 Introduction 25
3.2 Basic Concepts and Metrics 25
3.2.1 Entailment 25
3.2.2 Accuracy, Precision, Recall, and F1 26
3.2.3 Cosine Similarity 27
3.2.4 ROUGE 28
3.2.5 PageRank and TextRank 29
3.3 Techniques and Algorithms 29
3.3.1 Word Embeddings 29
3.3.2 A Centroid-Based Summarization Algorithm 30
3.3.3 The KMeans Algorithm 31
3.4 BERT Models 32
3.4.1 The Base BERT Model 32
3.4.2 The SBERT Model 33

4 Student Surveys Data **34**
4.1 Basic Concepts Related to Student Surveys 34
4.2 Student Surveys Scenarios 34
4.3 Student Surveys up to 2019 36
4.4 Student Surveys in 2019 37
4.5 Student Surveys in 2020/2021 38

5 Sentiment Analysis of Student Survey Comments **40**
5.1 Introduction 40
5.2 A Dictionary-based Approach 40
5.2.1 Description of the Dictionary-based Approach and the *CourseObservatory* Tool 41
5.2.2 Experiments and Results 42
5.2.3 Applications of the Results 44

5.3 A Neural Model Approach 46
5.3.1 Overview of the BERT Polarity Classification Model 46
5.3.2 Pre-Training Step 47
5.3.3 Training Step 47
5.3.4 Predictions 51
5.4 Chapter summary 53

6 Towards Comment Summarization 56
6.1 Introduction 56
6.2 Use of the Course Survey Data 59
6.3 Trending Topics Approaches 62
6.3.1 Market Basket Analysis 62
6.3.2 Topic Modeling 64
6.3.3 Lessons Learned from the Trending Topics Approaches 65
6.4 Partitioning Approaches 66
6.4.1 Clustering combined with the Centroid-based Summarization Algorithm 66
6.4.2 Attribute Partitioning 67
6.4.3 Lessons Learned from the Partitioning Approaches 70
6.5 Ranking Approach 71
6.5.1 The TextRank Algorithm Revisited 71
6.5.2 Top-k TextRank 71
6.5.3 TextRank combined with Clustering and the Centroid-based Summarization Algorithm 73
6.5.4 Lessons Learned from the Ranking Approaches 74
6.6 Entailment Approach 75
6.6.1 Computing entailment with BERT 75
6.6.2 Entailment combined with TextRank, Clustering, and the Centroid-based Summarization Algorithm 76
6.6.3 Lessons Learned with the Entailment Approach 77
6.7 Evaluation of the Comment Summarization Strategies 77
6.8 Further Experiments 80
6.9 Chapter summary 83

7 Conclusions 85

8 Bibliography 88

A Questionnaire for In-Person Disciplines 94

B Questionnaire for Online Disciplines 96

C Questionnaire for the Covid-19 Period 98

D Entailment 100

List of figures

Figure 3.1 Confusion matrix. 27
Figure 3.2 BERT representation text (Figure from Devlin et al. (2019)). 33

Figure 5.1 Steps of the CourseObservatory tool. 42
Figure 5.2 Comparison between average evaluation and comment sentiment classification of one teacher. 45
Figure 5.3 Distribution of comments by period and final status. 46
Figure 5.4 Distribution of the average score of all questions of a questionnaire from 2018. 48
Figure 5.5 Accuracy of the comments that coincide in sentiment for the manual and automatic classification of Table 5.3 49
Figure 5.6 Accuracy for *From scratch* and *Fine-tuned* using train set of 40, 80, 160, 320 and 640 comments. 51
Figure 5.7 F1 for *From scratch* and *Fine-tuned* using train set of 40, 80, 160, 320 and 640 comments. 52
Figure 5.8 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). 53
Figure 5.9 Distribution of the final classification of the comments from all surveys, using the fine-tuned model (shown in blue), and the classification of the comments from all surveys, using the score of Question O (shown in orange) for the courses Desing, Law, and Industrial Engineering. 54
Figure 5.10 Distribution of the final classification of the comments from all surveys, using the fine-tuned model (shown in blue), and the classification of the comments from all surveys, using the average score of all questions of a questionnaire (shown in orange) for the semester until 2019 for all university. 55

Figure 6.1 Summarization strategies. 59
Figure 6.2 Frequent terms for a topic. Each scale is different because the number of comments analyzed is different due to the filter applied. 63
Figure 6.3 Keyword and weight with which it contributes to the topic. 65
Figure 6.4 Optimal k for the data obtaining with the Elbow Method. 74
Figure 6.5 F1 measures for ROUGE-1, ROUGE-2, and ROUGE-L between the strategy summaries and the manual reference summaries. 78
Figure 6.6 Recommended comment summarization strategy. 80

List of tables

Table 4.1 Definitions 35
Table 4.2 Distribution of departments at PUC-Rio 36
Table 4.3 Data up to 2019 37
Table 4.4 2019 Data 38
Table 4.5 2020/2021 Data 39

Table 5.1 Distribution of the classified comments 43
Table 5.2 Results of the comparision between the tools 44
Table 5.3 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). 48
Table 5.4 Results of the setups 50

Table 6.1 Comments selected for the experiments. 60
Table 6.2 Sentences obtained from the selected comments. 60
Table 6.3 Comment topics using Market Basket Analysis 64
Table 6.4 Comments of the teacher for the experiments. 68
Table 6.5 TextRank Experiment Results 72
Table 6.6 TextRank Clusters and Centroid sentences 74
Table 6.7 Results for Entailment Strategy 77
Table 6.8 Reference and Strategy Summaries. 78

List of codes

List of Abreviations

BERT – Bidirectional Encoder Representations from Transformers

CBSA - Centroid-Based Summarization Algorithm

ETL – Extraction, Transformation and Loading

HEI – Higher Education Institution

LSTM – Long Short-Term Memory

NLP – Natural Language Processing

PDI – Pentaho Data Integration

ROUGE – Recall-Oriented Understudy for Gisting Evaluation

SA – Sentiment Analysis

SBERT – Sentence-BERT

STS – Sentence Textual Similarity

TE – Textual Entailment

Happiness can be found, even in the darkest of times, if one only remembers to turn on the light.

J. K. Rowling, *Harry Potter and the Prisoner of Azkaban.*

1
Introduction

1.1
Context and Motivation

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, HEIs must 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 teacher 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 thesis addresses these two challenges in the context of a set of questionnaires designed to assess teacher performance from (anonymized) student surveys applied at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio) from 2005 to 2021.

1.2
Problems Addressed and Contributions

This thesis addresses three problems, intuitively defined as follows:

Comment sentiment polarity problem. Given a set C of comments, classify the *sentiment polarity* of each comment in C.

Comment topic trending problem. Given a set C of comments, find a set of *topics* T that describe the comments in C.

Comment summarization problem. Given a set C of comments, find a subset $S \subset C$ such that S is much smaller than C and S conveys approximately the same meaning as C.

The thesis proposes two approaches to evaluate the sentiment polarity of students' comments. The first approach, described in Section 5.2, is based

on a manually created dictionary that lists terms that represent the sentiment to be detected in the students' comments. This approach was implemented as a tool, called *CourseObservatory*, which classifies the sentiment polarity of a set of students' comments and helps answer a set of questions that course coordinators (or department directors) may find useful. The tool was implemented in 2018 and tested with (anonymized) data from the surveys conducted from the second semester of 2005 to the second semester of 2018. The results were published in Jiménez et al. (2019) and indicated that the *CourseObservatory* tool outperforms a baseline tool.

The second approach, covered in Section 5.3, is based on the BERT language model, summarized in Section 3.4.1, and does not depend on a manually created dictionary. The BERT model was implemented using KERAS, running on GPUs (see Appendix **??**). The model was tested with data from the surveys conducted from the first semester of 2019 to the second semester of 2021. Studying this particular period is interesting because, in early 2020, the COVID-19 pandemic forced PUC-Rio to move all classes online, taught with the help of a videoconferencing software and a Learning Management System, and they so remained throughout 2020 and 2021. This change in instructional model offered 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 and throughout 2021 (late-COVID scenario). The results with the BERT model were published in Jiménez et al. (2021) and indicated that the model achieved very good performance, even when the set of manually annotated comments, used to train the BERT model, is small.

The thesis proposes two strategies, described in Section 6.3, to detect the trending topics of a set of student's comments: Market-Basket Analysis and Topic Modeling. These strategies were inspired by tweet trending topics summarization techniques, but they explore the specific context of students' comments.

As for the third problem, the thesis explores three approaches to summarize a set of comments, used in isolation or different combinations: *partitioning*, *ranking*, and *entailment*.

The intuition behind the *partitioning approach* is that the comments in a partition C_i should be redundant, that is, they should convey approximately the same meaning. Since it is difficult to argue that a set of comments is redundant, the partitioning step was implemented using essentially syntactical strategies. Section 6.4.1 investigates *clustering*, that is, grouping comments

by applying a clustering algorithm, based on a comment similarity measure, and Section 6.4.2 explores *attribute partitioning*, that is, grouping comments that have the same values for one or more attributes. In both cases, the summarization step was implemented using the centroid-based summarization algorithm, described in Section 3.3.2.

The intuition for adopting the *ranking approach* is that the top-ranked comments should be the most important ones. Comment ranking was implemented using TextRank, introduced in Section 3.2.5. The direct application of TextRank, discussed in Section 6.5.2, proved not to be adequate, so it was combined with clustering and the centroid-based summarization algorithm in Section 6.5.3.

The intuition for the *entailment approach* is that a comment c_i summarizes all comments c_j such that c_j transitively entails (i.e., implies) c_i. Entailment was implemented using a specially trained BERT model, described in Section 6.6.1. Again, the direct application of entailment proved not to be adequate. Section 6.6.2 then discusses how to use entailment in combination with TextRank, clustering and the centroid-based summarization algorithm.

Finally, several metrics have been proposed to evaluate text summarization strategies, such as the Rouge and Bleu metrics and n-gram novelty, as reviewed in Section 2.6. The thesis adopts the Rouge metric, with the provisos raised in Section 6.7, to compare the various strategies proposed to summarize a set of comments.

1.3 Structure

The thesis is organized as follows. Chapter 2 reviews related work in the different areas the thesis covers. Chapter 3 introduces concepts, metrics, models, and algorithms that will be used in the rest of the thesis. Chapter 4 presents the scenarios and definitions of the datasets used in the thesis. Chapter 5 describes the two approaches for sentiment analysis, with the experiments and results obtained. Chapter 6 covers the approaches, experiments, and results proposed to address the topic trending and summarization problems. Finally, Chapter 7 presents the conclusions of the results obtained during this investigation.

2
Related Work

2.1
Sentiment Analysis

Due to the explosion of social media, today the average citizen is overwhelmed by a large volume of text, available in digital format, expressing opinions, such as blogs, forum discussions, Twitter, among others. 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 (LIU, 2012; LIU; ZHANG, 2012). According to Pang & Lee (2008), 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.

Chaturvedia et al. (2018) 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. (2019) 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, Khoshgoftaar & Dittman (2015) also concentrated on Twitter data. They analyzed the impact of ten filter-based feature selection techniques on the performance of four classifiers.

Nazare et al. (2018) 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 & Qiu (2017) 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 & Srivastava (2018) 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 & Johnkhan (2018) analyzed 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, Ekbal & Bhattacharyya (2016) used classification algorithms, like Conditional Random Field (CRF) and Support Vector Machine (SVM), to classify comments from different Indian Web sites.

2.2 Sentiment Analysis in Higher Education

Zhou & Ye (2020) 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, Rita & Guerreiro (2018) 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.

Balahadia, Fernando & Juanatas (2016) presented a tool for the analysis of comments made by students to help improve the performance of teachers. The tool evaluates both quantitative and qualitative information. However, the tool is not directly applicable to the investigation in this article since it is limited to English.

Sindhu et al. (2019) 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.

Menaha et al. (2017) proposed a system based on the analysis of the repetition of keywords in a comment that extracts the main topic to which the comment refers. Once the topic is identified, they carry out a clustering process to classify the comments into positive or negative.

We have two publications in this area, previously created a tool for the analysis of student comments (JIMéNEZ et al., 2019) but it was limited to a fixed, manually created dictionary, which might therefore not take into account some relevant words. In Jiménez et al. (2021) we developed a NLP solution to classify the comments

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 & Kirupananda (2017) proposed an aspect-based sentiment analysis tool to evaluate the reputation of universities in Sri Lanka from users' comments on Facebook and Twitter, using the `StanfordCoreNLP` library to perform sentiment analysis. Lytras et al. (2016) 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 & Reddy (2017) 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.

2.3 Sentiment Analysis in Portuguese

Oliveira & Merschmann (2021) analyzed the combination of NLP preprocessing tasks (tokenization, POS tagging, stemming, among others) with three classifiers (Random 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

Souza & Vieira (2012) and Souza, Pereira & Dalip (2017) applied sentiment analysis to tweets. Although tweets and comments have similarities, at least in terms of size, our study focuses on students' comments in the context of course surveys, unlike these references.

2.4 BERT

One of the focuses of this research is identifying students' sentiments expressed in comments about teacher performance in Higher Education. It uses the pre-trained model called Bidirectional Encoder Representations from Transformers (BERT) (DEVLIN et al., 2019) 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 (DEVLIN et al., 2019). 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.

Other research also uses BERT to find feelings such as Pota et al. (2021) where they present an effective BERT-Based to find Sentiment Analysis in Twitter. Its purpose is to transform emojis and emoticons into plain text, which are one of the most used elements in tweets and then classified with BERT. Other sentiment analysis works also use BERT to solve this problem, such as (SUN; HUANG; QIU, 2019), (GAO et al., 2019) and (LI et al., 2020). These works serve as a guide but we cannot use them in their entirety, since they address text in languages other than Portuguese.

2.5 Semantics beyond individual sentences

In NLP, there are different research areas and, although they are closely intertwined with each other, they can be subdivided into categories. One of these categories is *semantics beyond individual sentences* and, within it, we find *sentence textual similarity* (STS) and *recognizing textual entailment* (TE), discussed in what follows.

2.5.1 Sentence Textual Similarity

Consider first the problem of sentence textual similarity. The applications of the STS can be found in different fields and it is that the possibility of

measuring the similarity between words, sentences, paragraphs, and documents plays an important role in computational linguistics.

Majumder et al. (2016) explained different measures for STS divided into categories: Topological/Knowledge-based, Statistical/Corpus Based and String based. Also for these categories they present tools and applications.

Reimers & Gurevych (2019) created a tool, based on BERT and called SBERT, to find sentence similarity and compared it with cosine similarity. They reduced considerably the effort to find the similarity between pairs of sentences using BERT and RoBERTa, without losing accuracy. SBERT can be used to compute sentence/text embeddings for more than 100 languages.

2.5.2 Recognizing Textual Entailment

Consider now the Textual Entailment (TE), we know that focuses on a directional relation between text fragments. TE considers a test sentence (entailing) and a hypothesis sentence (entailed) so that, intuitively, if a person reads the test sentence, then he/she would infer that the hypothesis sentence is most likely true. The relation is directional as analyzed by Tatar et al. (2009).

Unlike other textual entailment methods, Blake (2007) included the syntax and semantics to detect entailment. Their focus was on measuring the impact that sentence structure had on finding entailment. They developed two decision rules that each use features from a typed dependency grammar representation of the hypothesis and test sentences.

The number of datasets prepared for TE in different languages is gradually growing, but it is still not enough. Abdiansah, Azhari & Sari (2018) created a model to extract data from the Web to serve as a dataset for TE systems. Although the model was created for Indonesian, it can also be applied to other languages with some adjustments. Portuguese is one of the languages that does not yet have a large number of datasets for TE. Rocha and Lopes discussed these challenges and presented several approaches to address the task of recognizing entailment and paraphrases of a text written in Portuguese. ASSIN (Avaliação de Similaridade Semântica e INferência textual) is a dataset with semantic similarity scores and entailment annotations in Portuguese (REAL; FONSECA; OLIVEIRA, 2020). Some of published results on ASSIN can be find at http://nilc.icmc.usp.br/assin/.

2.6 Summarization

Text summarization received considerable attention recently. Allahyari et al. (2017) presented various extractive approaches for single and multi-document summarization. They explained the logic of the different methods, such as topic representation, frequency-driven methods, graph-based and machine learning techniques. Lemberger (2020) also analyzed several strategies and described three models that use Deep Learning and therefore implement a purely statistical approach to the summarization task. He concluded that these models work well for short documents but is not clear if they can be used for large documents.

Kleindessner, Awasthi & Morgenstern (2019) presented a strategy for summarization where they choose k prototypes to summarize a dataset. They consider the grouping of k centers under an equity constraint, motivated by the application of groupings based on centroids. They presented a 5-approximation algorithm for two groups. For more than two groups, they try an upper bound on the approximation factor that increases exponentially with the number of groups. But they were unable to answer whether this exponential dependence is necessary or whether the analysis is imprecise.

Rossiello, Basile & Semeraro (2017) proposed a method based on centroids to summarize a text that tried to solve the deficiency of the use of bag-of-words to capture semantic relationships between concepts, when strongly related sentences are compared.

Usually, summarization techniques are not directly applied to tweets, since tweets are written by different users and tweets are short sentences. Rather, trending topic analysis techniques are applied to summarize tweets. To correct for dialect bias, Naik et al. (2018) employed a framework that takes an existing text summary algorithm as a black box and, using a small set of sentences with various dialects, returns a summary that is relatively more diverse. Tweets are sent for semantic analysis, weights are assigned to the tweets, and a graph is formed for clustering. Similar tweets are clustered using the Particle Swarm Optimization algorithm. Finally, a tweet from each cluster is chosen to be included in the summary. This approach was considered in our investigation.

Finally, the common metric to evaluate the quality of a summary is the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (LIN, 2004). ROUGE is a set of metrics, rather than just one. Alternative metrics and improvements to ROUGE were proposed by Steinberger & Jezek (2009), Bhandari et al. (2020), Fabbri et al. (2021).

3
Background

3.1
Introduction

This chapter collects background material to facilitate the reading of chapters 5 and 6. It is intended neither to be a tutorial nor a survey of the topics covered, which can be easily found in the literature.

Since the topics covered are quite varied, they are organized in three sections. Section 3.2 contains definitions of concepts and metrics used in chapters 5 and 6. Section 3.3 covers word embedding, a centroid-based summarization algorithm, and the KMeans algorithm, used in Chapter 6. Finally, Section 3.4 summarizes the basics of the BERT neural network models used in chapters 5 and 6.

3.2
Basic Concepts and Metrics

3.2.1
Entailment

Intuitively, we say that a sentence A *entails* a sentence B when A implies B, as the following examples illustrate (BENTIVOGLI et al., 2016):

- Example 1
  - *A:* A man is talking to a woman.
  - *B:* A man and a woman are speaking.
- Example 2
  - *A:* Two children and an adult are standing next to a tree limb.
  - *B:* Three people are standing next to a tree limb.
- Example 3
  - *A:* A man and two women in a darkened room are sitting at a table with candle.
  - *B:* The group of people is sitting in a room which is dim.

Naturally, if a sentence A entails a sentence B, it does not necessarily mean that B entails A. Entailment is used, for example, to derive answers

from stored information in a question answering system, and to summarize documents by filtering sentences that do not include new information.

Some systems define three categories of entailment between two sentences, A and B:

- *positive entailment*, when A implies B
- *negative entailment*, when A refutes B
- *neutral entailment*, when A and B have no correlation

while others consider just two categories:

- *entailment*, when A implies B
- *none*, otherwise

3.2.2 Accuracy, Precision, Recall, and F1

Many performance metrics have been proposed to evaluate a classification model, such as accuracy, precision, recall, F1-score, each having advantages and disadvantages.

Most of these metrics are defined based on the confusion matrix for two classes, shown in Figure 3.1. This matrix has two rows and two columns that indicate the number of *false positives, false negatives, true positives*, and *true negatives*, with the following meaning:

- **True Positive (TP):** The predicted class was positive and the actual class was also positive.
- **False Positive (FP):** The model predicted positive and the actual class was negative.
- **False Negative (FN):** The predicted class was negative and the actual class was positive.
- **True Negative (TN):** The predicted class was negative and the actual class was also negative.

Based on the confusion matrix, the most common metrics are defined as follows:

Accuracy - Is a ratio of correctly predicted observations to the total number of observations. A good model has high accuracy.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3-1}$$

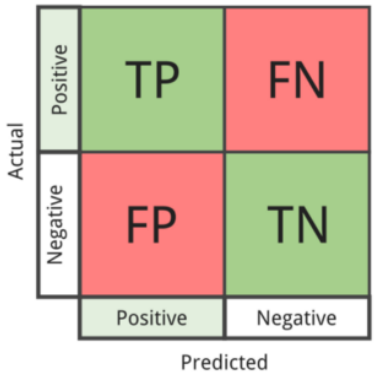

Figure 3.1: Confusion matrix.

Precision - Is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision means that more relevant than irrelevant results are returned

$$Precision = \frac{TP}{TP + FP} \tag{3-2}$$

Recall - Is the ratio of correctly predicted positive observations to all observations in the actual class. A high recall means that most of the relevant results are returned.

$$Recall = \frac{TP}{TP + FN} \tag{3-3}$$

F1 - Is the weighted average of Precision and Recall. F1 is useful if you have an uneven class distribution.

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{3-4}$$

3.2.3 Cosine Similarity

Cosine similarity is a measure of the similarity between two vectors. Recall that the value of the cosine of the angle between two vectors is in the closed interval [-1,1]; it is 1 when the angle is 0 (both vectors point in the same direction); and -1, when the angle is 180^o (the vectors point in opposite directions). The cosine similarity between two vectors A and B is defined as:

$$similarity(A, B) = \frac{A * B}{\|A\| * \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3-5}$$

where A_i and B_i are the components of vectors A and B, respectively.

Python's `sklearn` (PEDREGOSA et al., 2011) library has a function, `cosine_similarity(A, B)`, which computes the cosine similarity between all vectors in **A** and all vectors in **B**. This function returns a matrix with the cosine similarity scores for all possible such pairs.

3.2.4 ROUGE

ROUGE (LIN, 2004) is a set of metrics to compare a model-generated text and a "reference" text, usually manually created. This thesis uses ROUGE-N and ROUGE-L, defined as follows:

- ROUGE-N measures the number of matching n-grams between the model and the reference. An *n-gram* is a grouping of tokens/words; a unigram (1-gram) consists of a single word, a bigram (2-gram) of two consecutive words, and so on.
- ROUGE-L measures the longest common subsequence (LCS) between the model and the reference – the longer the shared sequence is, the more similar the model and the reference are.

ROUGE may be combined with the recall, precision, and F1 scores as follows:

Recall - defined as the number of overlapping n-grams found in both the model and reference, divided by the total number of n-grams in the reference.

$$Recall = \frac{count_{match}(gram_n)}{count_{reference}(gram_n)} \tag{3-6}$$

Precision - defined as the number of overlapping n-grams found in both the model and reference, divided by the total number of n-grams in the model.

$$Precision = \frac{count_{match}(gram_n)}{count_{model}(gram_n)} \tag{3-7}$$

F1 - is a measure of the model performance that relies not only on the model capturing as many words as possible (recall), without outputting irrelevant words (precision).

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{3-8}$$

Python offers the `rouge` library for computing these metrics. The *get_scores(model, reference)* function computes ROUGE-1, ROUGE-2 and ROUGE-L by default and returns the recall, precision, and F1 for each metric.

3.2.5 PageRank and TextRank

PageRank defines a relevance score for Web pages, which is then used to rank the pages search engines return to the users (LESKOVEC; RAJARAMAN; ULLMAN, 2020).

Consider a square matrix M with n columns and rows, where n is the number of Web pages, defined as follows:

$$M[i][j] = \begin{cases} 1/k, & \text{if page } j \text{ has } k \text{ out-links and one of them is to page } i \\ 0, & \text{otherwise} \end{cases}$$

The PageRank score is recursively defined as follows:

$$v_{i+1} = d.M.v_i + (1-d)\mathbf{e}/n \tag{3-9}$$

where

- v_i is a vector such that $v_i[p]$ is the PageRank estimation of page p at iteration i
- d is a damping factor, usually in the range 0.8 to 0.9
- $\mathbf{e}$ is the unit vector of size n

Inspired on PageRank, the *TextRank* (MIHALCEA; TARAU, 2004) score was defined to rank natural language sentences. Instead of a Web page, TextRank uses sentences and substitutes the Web page transition probability for the similarity between two sentences. The similarity is also represented as a square matrix M, as for PageRank, so that Eq. 3-9 can be used to define TextRank.

3.3 Techniques and Algorithms

3.3.1 Word Embeddings

Word embedding is a NLP technique where words or phrases are represented as vectors of real numbers. Gilyadov (2017) explains that there are two methods to compute word embeddings: *count-based methods*, and *predictive methods*. Both methods assume the Distributional Hypothesis, which states that linguistic items with similar distributions have similar meanings, more simply, words that appear in the same contexts share semantic meaning.

Python has several libraries to compute word embeddings. We chose *Gensim* (ŘEHŮŘEK; SOJKA, 2010) because it can easily process large and

Web-scale corpora by using its incremental online training algorithms, it is robust, it has been in use in various systems, and it provides efficient multicore implementations of various popular algorithms. In our application, *Gensim* would return a list where each word of the sentences is represented as a vector.

3.3.2 A Centroid-Based Summarization Algorithm

The Centroid-Based Algorithm is a text summarization algorithm proposed by Rossiello, Basile & Semeraro (2017), which has the following major steps, detailed in the rest of this section:

1. Build a centroid vector, using word embeddings.
2. Select the meaningful words that occur in the text.
3. Compute a vector embedding representation for each sentence by summing the vector embeddings of the words that occur in the sentence.
4. Compute the cosine similarity between the centroid vector and the embedding representation of each sentence.
5. Select the sentences with the highest value of cosine to return as summary of the text.

In what follows, we explain these steps in more detail.

The first step is to build the word-embedding of the text, for this we will use the library described in Section 3.3.1. In this case, the algorithm also depends on computing the *tf (term frequency)* and the *idf (inverse document frequency)* of each word that occurs in the sentences. Intuitively, the *term frequency* is the number of occurrences of a term in a document, divided by the total number of term occurrences in the document. The *inverse document frequency* is a measure of how much information a term provides, i.e., if it is common or rare across all documents. These two measures are defined as follows:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{u \in d} f_{u,d}} \tag{3-10}$$

$$idf(t,D) = log\frac{N}{1+N_t} \tag{3-11}$$

where

- d is a document in a set of documents D
- t is a term
- $f_{t,d}$ is the number of times t occurs in d

– N is the total number of documents in D

– N_t is the total number of documents in D where t occurs (the denominator of the definition of *idf* is adjusted to $1 + N_t$ because t may not occur in any document in D)

The Python library `sklearn` implements these measures.

Assume that the *tf* and *idf* of each word in each sentence have been computed. To create a centroid vector, we select those words having a $tf.idf$ greater than a *topic threshold*. Rossiello, Basile & Semeraro (2017) suggested to select a value between [0.3, 0.5] for the topic threshold, based on several experiments.

The next step is to compute an embedding representation for each sentence by summing the vectors of each word in the sentence.

To compute the similarity between each sentence and the centroid vector, Rossiello, Basile & Semeraro (2017) proposed to use cosine similarity (see Section 3.2.3). The result is represented as a vector. To select the most representative sentences of the group, just choose those with the highest similarity values in this matrix.

The code of this algorithm is in GitHub[1].

3.3.3 The KMeans Algorithm

KMeans is a clustering algorithm (LESKOVEC; RAJARAMAN; ULLMAN, 2020) that requires specifying the number of clusters and assumes an Euclidean space to minimize the within-cluster sum of squares. Let

– S_i be a cluster

– x_j be the observations (encoded sentences) in S_i

– μ_i be the centroid of S_i

Then, the *within-cluster sum of squares* for cluster S is defined as

$$argmin_{S_i} \sum_{i=1}^{h} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \tag{3-12}$$

To find which is the most suitable number of clusters for the data, there are several strategies, including the *Elbow Method*, which is based on identifying the sum of the squared distances of the samples from the center of the closest cluster. As the number of clusters increases, the sum of the squared distance tends to zero; the graphical representation of the function that represents this

[1] Available at https://github.com/hguillot/centroid_based_summarization_algorithm

sum looks like an “arm”, and the value of the point where the “elbow” of this “arm” is located will be the optimal number of clusters.

Python’s `sklearn` library has an implementation of the KMeans algorithm and other functions to help analyze the “elbow” value.

3.4 BERT Models

There are different versions of BERT (KHAN, 2019), selected according to the data volume and the analyzes that will be carried out. Chapter 5 analyses students’ comments, which are simple sentences, and adopts the base BERT (DEVLIN et al., 2019). Chapter 6 addresses comment summarization and resorts to SBERT (REIMERS; GUREVYCH, 2019), a modification of BERT that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity.

3.4.1 The Base BERT Model

BERT (DEVLIN et al., 2019) is a neural network model that achieves outstanding results on several NLP (Natural Language Processing) tasks, including entity recognition and sentiment analysis. BERT’s architecture is based on multi-layered transformers, which are particularly optimized to be trained using GPUs and TPUs with significant amounts of data.

BERT has two types of architectures that differ in four fundamental aspects: the number of hidden layers, the number of attention heads, the hidden size of the feed-forward networks, and the maximum sequence length parameter. The architectures are:

– *base*, with 12-layer, 768-hidden, 12-heads, 110M parameters
– *large*, with 24-layer, 1024-hidden, 16-heads, 340M parameters

The choice of the architecture depends on the data to be used.

To implement the neural network models used in Chapter 6, we adopted *Keras*[2], a deep learning API written in Python. *Keras* is simple, flexible and powerful, and was developed with a focus on enabling fast experimentation. The way to create a model with *Keras* is quite intuitive: it is enough to declare the input layer, the output layer, and then the hidden layers that are necessary.

One of the most important steps of BERT is pre-training. This step allows the model to learn language patterns that are independent of the task in question, and does not require large amounts of annotated data.

[2]https://keras.io

BERT models are pre-trained with large datasets (in the order of millions of documents) on general tasks, such as masked language models or next sentence predictions (DEVLIN et al., 2019).

BERT extracts word and sentence embedding vectors from text data, which are used as inputs to downstream models. BERT uses two special tokens: `[CLS]`, which is added in front of every input example; and `[SEP]`, which separates text segments. Figure 3.2 illustrates how BERT represents text inputs.

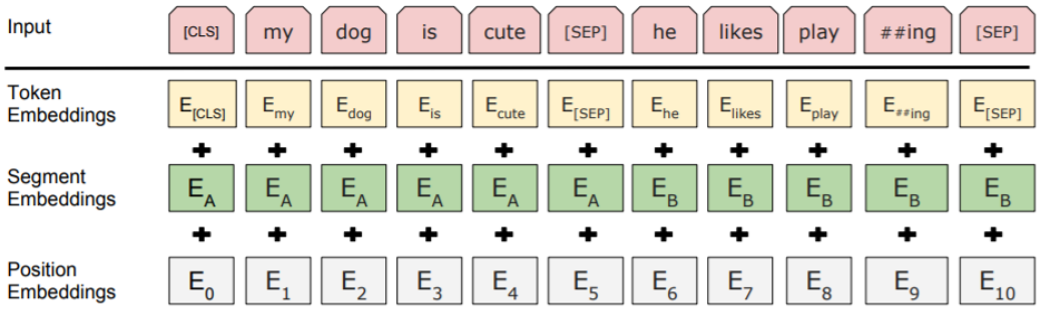

Figure 3.2: BERT representation text (Figure from Devlin et al. (2019)).

After the pre-training step, the models go through a training step, as discussed in Section 5.3.3.

3.4.2 The SBERT Model

SBERT (REIMERS; GUREVYCH, 2019) is a Python framework, based on PyTorch and Transformers, to encode sentences and analyze their similarity.

As for BERT, the first step to use SBERT is to create a sentence transformer model that maps sentences/text to embeddings.

The second step is to encode the sentences, using the *encode* function that SBERT provides. This function creates an array with the embeddings corresponding to each sentence in the dataset.

With the sentences encoded, the next step is to compute the similarity between sentences and cluster the set of sentences. The creators of SBERT recommend using cosine similarity, described in Section 3.2.3.

4
Student Surveys Data

4.1
Basic Concepts Related to Student Surveys

Throughout this thesis, 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. Depending on the demand, a course may have more than one class, not necessarily taught by the same teacher. Among the types of courses offered, one finds monographs, seminars, projects, and regular classroom courses. Finally, 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.

Table 4.1 contains a complete list of the concepts used.

4.2
Student Surveys Scenarios

The Pontifical Catholic University of Rio de Janeiro - PUC-Rio is a private university in Rio de Janeiro, Brazil. It is structured in 4 centers: Center of Theology and Human Sciences (CTCH), Center of Social Sciences (CCS), Technical-Scientific Center (CTC), and Center of Biological and Health Sciences (CCBS). Each center is subdivided in departments, summarized in Table 4.2 (as of 2021).

Since 2005, at the end of each semester, the Central Coordination of Planning and Evaluation (CCPA for its acronym in Portuguese) applies an online evaluation where students evaluate each of the classes s/he took in that semester. The evaluation is voluntary and completely anonymous and, to preserve anonymity, only classes with more than 3 students enrolled are evaluated. Recalling Article 1 of Resolution No. 510 of April 7, 2016, that exempts "research with databases, whose information is aggregated, without the possibility of individual identification", the work reported in this thesis needs no formal approval by a Research Ethics Chamber.

The typical case is a class taught by one teacher, in which case the student receives a questionnaire with questions about this single teacher. However, a class may be taught by more than one teacher, the situations being:

– 2 teachers, both with a teaching load: both are responsible for the class so it would be necessary to evaluate them both.

Table 4.1: Definitions

Definition	Description
# students	Number of students in the university.
# enrollment	Number of students enrollment in courses (if a student is enrolled in more than one course counts multiple times)
# enrollment participants	Number of students enrollment in courses that participated in the evaluation
# teachers	Number of teachers who taught courses
# teachers by courses	Number of teachers who taught different courses (if a teacher taught two different courses count twice, but if the same course gives two offers only count as one)
# teachers by classes	Number of teachers who taught different courses or several offers of the same courses (if a teacher taught two different courses or if the same course gives two offers counts twice)
# teachers evaluated	Number of teachers evaluated
# teachers evaluated by courses	Number of teachers evaluated taking into account the courses (if the teacher was evaluated by two different courses it counts as two teachers evaluated)
# teachers evaluated by classes	Number of teachers evaluated taking into account the courses and offers (if the teacher was evaluated by two different courses or if the teacher was evaluated in the same course but by two different offers counts as two evaluated teachers)
# teachers comments	Number of comments with any text for teachers
# empty teachers comments	Number of empty comments for teachers
# courses	Number of university courses
# classes	Number of courses and offers (if a course has two offers it counts as two)
# courses evaluated	Number of courses evaluated
# classes evaluated	Number of courses and offers evaluated (if a course has two offers it counts as two)
# courses comments	Number of comments with any text for the courses
# empty courses comments	Number of empty comments for the courses

- 2 teachers, but only one with a teaching load: the teacher with a teaching load will be evaluated, as he is the person in charge of the class. The other teacher is the coordinator and he is only registered as a teacher of that class to be aware of the development during the semester, but s/he will not be evaluated by the students.
- More than 2 teachers: the teacher with the highest teaching load will be evaluated quantitatively. The students will only be able to comment on the performance of the other teachers.

A commission composed of teachers from the different centers of the university and some members of the CCPA is responsible for the creation of the questionnaire. This questionnaire has to be objective and applicable to all

Table 4.2: Distribution of departments at PUC-Rio

Center	Department	Acronyms in Portuguese
CTCH	Architecture and Urbanism	DAU
	Arts and Design	DAD
	Education	EDU
	Philosophy	FIL
	Literature	LET
	Psychology	PSI
	Theology	TEO
CCS	Administration	ADM
	Social Sciences	SOC
	Social Communication	COM
	Law	JUR
	Economy	ECO
	Geography and Environment	GEO
	History	HIS
	Social Service	SER
	Institute of International Relations	IRI
CTC	Basic Cycle of the Scientific Technical Center	CBCTC
	Professional Cycle of Engineering	CCPE
	Civil and Environmental Engineering	CIV
	Electrical Engineering	ELE
	Industrial Engineering	IND
	Mechanical Engineering	MEC
	Chemical and Materials Engineering	DEQM
	Physics	FIS
	Informatics	INF
	Mathematics	MAT
	Chemistry	QUI
CCBS	Biology	BIO
	Medicine	MEDPUC

teachers and disciplines at the university. The questionnaire may sometimes be modified, as for example in the case of the COVID-19 in 2020.

The results of the evaluation are made available to the teachers, coordinators and directors of the university through an online Web site. The teachers can access only their results, the coordinators and directors of the departments view the results of all teachers of the department, the deans of the centers access the results of the teachers of the departments that make up the center, and finally the central coordinators have access to all the results.

4.3 Student Surveys up to 2019

Between 2005 and 2019, the survey questionnaire had 10 questions about the usefulness of the course, the bibliography adopted, and the pedagogical abilities of the teacher, among others. There was also a free text area where

students could voice their opinion.

A disadvantage of this questionnaire is that it was applied to a class regardless of whether it was taught by more than one teacher. Hence, comments could be about several teachers, opinions about the teacher, or opinions about the course, all intermixed.

Table 4.3 shows the distribution of the data, by semester.

Table 4.3: Data up to 2019

Period	#Courses	#Classes	#Teachers	#Students	#Comments
2005.2	941	1,930	1,058	2,241	7,735
2006.1	896	1,885	1,034	1,848	6,207
2006.2	911	1,878	1,026	1,915	6,716
2007.1	888	1,837	1,002	1,575	5,727
2007.2	865	1,633	952	1,126	4,071
2008.1	868	1,671	970	1,151	4,177
2008.2	986	1,986	1,037	1,970	6,603
2009.1	997	2,088	1,048	2,180	7,399
2009.2	1,042	2,365	1,088	2,326	8,202
2010.1	1,033	2,444	1,091	2,582	9,242
2010.2	1,002	2,116	1,060	2,045	6,791
2011.1	1,022	2,226	1,108	2,160	7,060
2011.2	1,040	2,216	1,077	2,008	6,642
2012.1	1,053	2,245	1,083	2,136	6,845
2012.2	1,026	2,096	1,056	1,715	5,551
2013.1	1,078	2,325	1,093	2,090	7,017
2013.2	1,066	2,241	1,064	2,089	6,776
2014.1	1,131	2,468	1,138	2,362	7,970
2014.2	1,062	2,191	1,059	1,694	5,603
2015.1	1,124	2,429	1,121	2,304	7,667
2015.2	1,140	2,344	1,101	1,946	6,389
2016.1	1,164	2,363	1,103	2,020	6,543
2016.2	1,075	2,081	1,019	1,594	5,069
2017.1	1,099	2,178	1,043	1,698	5,553
2017.2	1,092	2,104	1,020	1,539	4,958
2018.1	1,179	2,257	1,058	1,874	6,247
2018.2	998	1,827	949	1,213	3,951

4.4 Student Surveys in 2019

In 2019, the evaluation system has been restructured, modifying the questions and breaking the questionnaire into two: one for the teacher and another for the course. The questionnaire was also different for *in-person* and *online* classes (note that a course may have in-person classes and online classes on the same semester):

– in-person classes: composed of 17 questions about the teacher and 8 questions about the course (see Appendix A)
– online classes: composed of 15 questions about the teacher and 7 questions about the course (Appendix B)

A student gave a score, in the Likert scale (1-5), for each of the questions; there was also a free text area where the student could comment on the teacher or the course, depending on the questionnaire.

Table 4.4 shows the distribution of the data for 2019.1 and 2019.2.

Table 4.4: 2019 Data

	in-person classes		**online classes**	
	2019.1	**2019.2**	**2019.1**	**2019.2**
# students	11.160	10.301	806	797
# enrollment	63.460	57.868	838	833
# enrollment participants	32.758	12.953	448	170
# teachers	1.130	1.089	22	22
# teachers by courses	2.572	2.496	24	24
# teachers by classes	3.559	3.639	260	256
# teachers evaluated	1.108	1.052	22	22
# teachers evaluated by courses	2.255	2.126	24	24
# teachers evaluated by classes	2.930	2.764	202	126
# teachers comments	3.488	2.122	41	18
# empty teachers comments	31.537	12.113	532	210
# courses	1.421	1.937	11	10
# classes	2.971	2.782	150	147
# courses evaluated	1.336	1.283	11	10
# classes evaluated	2.642	2.392	119	76
# courses comments	1.689	989	25	10
# empty courses comments	31.069	11.964	423	160

4.5 Student Surveys in 2020/2021

With the arrival in Brazil of COVID-19, teachers faced the new challenge of teaching classes using videoconferencing applications and began digitizing course materials. It was also necessary to rework the questionnaire. This emergency questionnaire was composed of 21 questions about the teacher, 4 questions about the course, and 3 free text areas for positive comments, negative comments, and to suggest ideas for this new environment (see Appendix C).

Table 4.5 shows the distribution of the data for 2020.1, 2020.2, and 2020.1.

Table 4.5: 2020/2021 Data

	2020.1	**2020.2**	**2021.1**
# students	10.174	9.810	10.060
# enrollment	57.677	59.244	59.375
# enrollment participants	30.037	26.465	28,382
# teachers	1.071	1.057	1.052
# teachers by courses	2.282	2.155	2.232
# teachers by courses and offers	3.173	2.961	3.039
# teachers evaluated	1.061	1.054	1.049
# teachers evaluated by courses	2.249	2.137	2.207
# teachers evaluated by courses and offers	3.082	2.859	2.941
# teachers comments	3.866	2.649	3.184
# empty teachers comments	29.024	27.149	28.704
# courses	1.379	1.349	1.357
# courses by offers	2.906	2,733	2.824
# courses evaluated	1.285	1.249	1.285
# courses evaluated by offers	2.598	2.410	2.499
# courses comments	4.296	2.937	3.582
# empty courses comments	85.815	76.458	81.563

5
Sentiment Analysis of Student Survey Comments

5.1
Introduction

This chapter focuses on the *polarity classification* task, whose focus is to classify comments that 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.

The chapter describes two approaches to classify the polarity of students' comments into "positive", "negative", or "neutral". The first approach, described in Section 5.2, is based on a manually created dictionary that lists terms that represent the sentiment to be detected in the students' comments. This approach was implemented as a tool, called *CourseObservatory*, which classifies the polarity of a set of students' comments and helps answer a set of questions that course coordinators (or department directors) may find useful.

The second approach, covered in Section 5.3, is based on the BERT model, summarized in Section 3.4.1, and does not depend on a manually created dictionary. The BERT model was implemented using KERAS, running on GPUs.

The results reported in Section 5.2 were published in Jiménez et al. (2019), with data until the second semester of 2018. The results indicated that the *CourseObservatory* tool outperforms a baseline tool. The results described in Section 5.3 were published in Jiménez et al. (2021) and had as motivation to investigate how students reacted to the move to online classes forced by the COVID-19 pandemic, using data from 2019, 2020, and 2021. The results indicated that the second approach achieved very good performance, even when the set of manually annotated comments is small. Our corresponding code is available at GitHub[1].

5.2
A Dictionary-based Approach

This section explains the dictionary-based approach and the *CourseObservatory* tool. This is a fairly simple and naive approach, but it was worth

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

experimenting with to get an idea of how it worked. It compares the *CourseObservatory* tool with a baseline tool and shows how to use it to help answer some questions that department coordinators may have.

5.2.1 Description of the Dictionary-based Approach and the *CourseObservatory* Tool

The approach has several steps. The first step separates the students' comments using punctuation marks, such as "." and ",", and keywords, such as "porém" ("although") and "mesmo assim" ("even though"). The idea is to be able to analyze each student's idea separately, and not the comment as a whole, because relevant information could be lost when analyzing a comment as conveying a single idea.

The next step transforms comments into a canonical form by first converting all words to lower case and then eliminating stop words. For example, "não foi bom" ("it was not good") and "não é bom" ("it is not good") were both transformed to "não bom" ("not good").

Then, the user must construct a dictionary based on an analysis of the most common terms that occur in the comments. For the construction of this dictionary, the user can adopt any other tool he/she wishes to obtain a set of words that better represent the feelings to be analyzed in the comments. Also, the user must manually classify each term as positive or negative.

The final step uses the dictionary to separately analyze the sentiment of each idea expressed in a comment C. It groups the ideas in C to classify C as P - "positive", N - "negative", E - "mixed" (when the comment had both positive and negative phrases), and Ne - "unknown" (when it was not possible to classify the comment). One very important thing to keep in mind when classifying comments is the context in which the words are used. For example, it is not enough to use words like "boa" and "bom" ("good") to classify a comment as positive, because in the context "não bom" ("not good") the comment is negative.

The *CourseObservatory* tool implements these steps, using the Pentaho Data Integration (PDI) [2] suite, as illustrated in Figure 5.1. PDI is a platform for the Extraction, Transformation, and Loading (ETL) process, which allows transformations and jobs to be carried out in a very easy way. It has a very intuitive desktop application that allows the extraction of data from

[2]https://www.hitachivantara.com/en-us/products/data-management-analytics.html?source=pentaho-redirect

different sources and multiple database managers, and implements many of the transformation functions, facilitating the programming process.

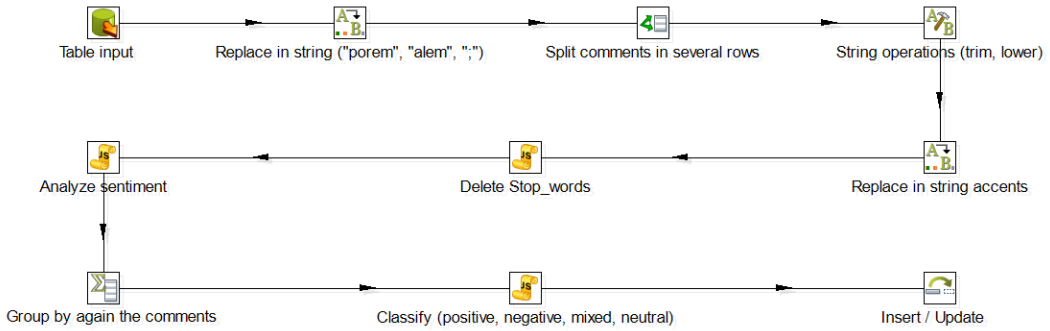

Figure 5.1: Steps of the CourseObservatory tool.

5.2.2 Experiments and Results

To experiment with the CourseObservatory tool, we used the comments from mid-2005 to the first semester of 2018, with a total of 918,439 records. After eliminating all blank comments, the dataset was left with 168,760 records. Table 4.3 shows the distribution of the records per semester.

For the construction of the dictionary, we applied an N-Gram Extraction Tool, based on the study of Lü (2004). Words from 1 to 5-gram, whose absolute frequency was larger than 100, were extracted and then analyzed to uncover which words reflect feelings to be placed in the dictionary. The final dictionary consisted of 211 terms, where 123 were manually classified as negative, 86 as positive, and 2 as mixed. Several terms were very similar because it was necessary to consider the teacher's gender ("professor", "professora"), spelling mistakes ("exelente" instead of "excelente"), and the young people's jargon ("mto" instead of "muito"). Using the dictionary, we then classified the comments using the *CourseObservatory* tool. Table 5.1 shows the distribution of the classification of the comments, by period, that the tool produced. Out of the 168,760 comments, 79% were classified and, for the classified comments, 56% were positive, 2% negative and 18% mixed.

To evaluate the performance of the *CourseObservatory* tool, we randomly chose 150 records, and manually classified them. Then, we compared the manual classification of these comments with the results of the tool and a Microsoft Excel (MEANINGCLOUD, 2016) tool that performs sentiment analysis. To use this component, it is necessary to install the Excel MeaningCloud, which has a sentiment analysis tool, limited to 10,000 records at a time. The tool returns the results of the analysis in a new tab, and classifies the comments into 6

Table 5.1: Distribution of the classified comments

Period	Positive	Negative	Mixed	Unknown	Total
2005.2	3,145	1,738	1,039	1,813	7,735
2006.1	2,700	1,252	795	1,460	6,207
2006.2	2,919	1,401	954	1,442	6,716
2007.1	2,452	1,256	884	1,135	5,727
2007.2	1,726	868	605	872	4,071
2008.1	1,742	917	609	909	4,177
2008.2	2,776	1,454	883	1,490	,6,603
2009.1	3,271	1,582	941	1,605	7,399
2009.2	3,515	1,645	1,057	1,985	8,202
2010.1	4,083	1,744	1,192	2,223	9,242
2010.2	3,112	1,356	876	1,447	6,791
2011.1	3,280	1,431	897	1,452	7,060
2011.2	2,985	1,340	970	1,347	6,642
2012.1	3,033	1,463	1,014	1,335	6,845
2012.2	2,584	1,096	782	1,089	5,551
2013.1	3,213	1,488	1,018	1,298	7,017
2013.2	3,130	1,374	1,023	1,249	6,776
2014.1	3,750	1,583	1,149	1,488	7,970
2014.2	2,529	1,092	785	1,197	5,603
2015.1	3,386	1,617	1,172	1,492	7,667
2015.2	2,981	1,200	930	1,278	6,389
2016.1	2,937	1,373	988	1,245	6,543
2016.2	2,271	1,073	710	1,015	5,069
2017.1	2,425	1,193	881	1,054	5,553
2017.2	2,233	989	703	1,033	4,958
2018.1	2,769	1,306	965	1,207	6,247
Total	74,947	34,831	23,822	35,160	168,760

categories: positive (P), negative (N), very positive (P+), very negative (N+), neutral (NEU), and none (NONE). The Excel tool performs sentiment analysis in several languages, including Portuguese (one of the reasons for choosing this tool as baseline). Note that the Excel tool returns 6 classifications (P, P+, N, N+, NONE, NEU), while the *CourseObservatory* tool returns 4 classifications (P, N, E, Ne); we then mapped P+ and P to P, N+ and N to N, NEU to E, and NONE to Ne. For the comparison, we adopted precision and recall, introduced in Section 3.2.2. Table 5.2 shows the results of the comparison, which indicates that the *CourseObservatory* tool outperforms the Excel tool by a large margin.

Table 5.2: Results of the comparision between the tools

		Excel	CourseObservatory
Precision	Positive	0.74	**0.93**
	Negative	0.51	**0.90**
	Mixed	0.43	**0.77**
	Average	0.56	**0.87**
Recall	Positive	0.69	**0.83**
	Negative	0.62	**0.76**
	Mixed	0.28	**0.83**
	Average	0.53	**0.81**

5.2.3 Applications of the Results

This section shows how to use the CourseObservatory tool to help answer some questions that department coordinators may have.

Consider the first question:

Q1. Is there a dependency between the average teacher evaluation and the tendency of the comments of his/her students?

To help answer this question, one may compare the average evaluation obtained by a given teacher per semester, computed from the class evaluation questionnaire, with the distribution of the classification of the students' comments, obtained with the CourseObservatory tool. It must be pointed out that the participation of students in the course surveys is not mandatory, but for a course or teacher to be evaluated, he must have been evaluated by a representative percentage of the total number of students enrolled.

For example, consider Figure 5.2, which shows data for a given teacher (omitted for privacy) over several semesters. The top part of the figure shows the average evaluation and the bottom part shows the number of positive, negative, and neutral comments, as classified by the CourseObservatory tool. Observe from the bottom part that he/she is a teacher who usually receives more than ten comments per semester and, from the top part of the figure, that his/her semester average evaluation is usually above 80. In the three semesters his/her evaluation was below 75 points – the periods 2005.2, 2006.1, and 2017.1 – in two of them he/she received a considerable number of negative comments. But this is not always the case. In the periods 2010.1 and 2012.1, where he/she obtained an evaluation of 85 points, he/she also received more negative comments than positive ones.

Therefore, it cannot be concluded that there is a relationship between the average teacher evaluation and the tendency of the comments of the

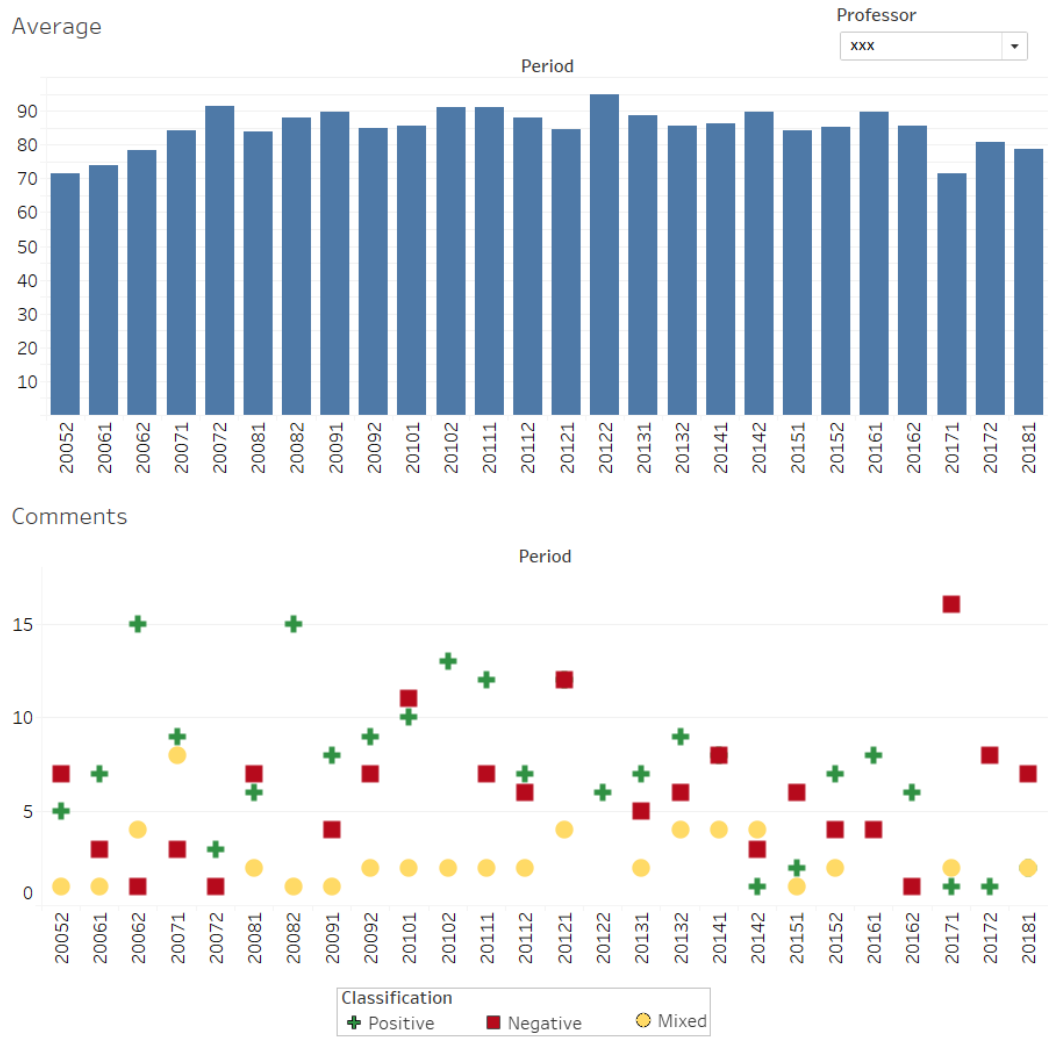

Figure 5.2: Comparison between average evaluation and comment sentiment classification of one teacher.

students. This fact indicates to the coordinators the importance of reviewing all comments, including the comments about teachers with a good average evaluation, as they can have a large number of negative comments, as Figure 5.2 illustrates.

Consider now the second question:

Q2. Can the sentiment tendency of a student's comment be influenced by the final status ("approved"/"failed") achieved in the course?

It should be remarked that the students receive their final status before evaluating the teacher.

Figure 5.3 shows the classification of the comments of the students according to their final status in each course, including all comments by period. Note that most of the comments in a period are made by students who were approved and the majority of the comments are positive. Also observe that, by

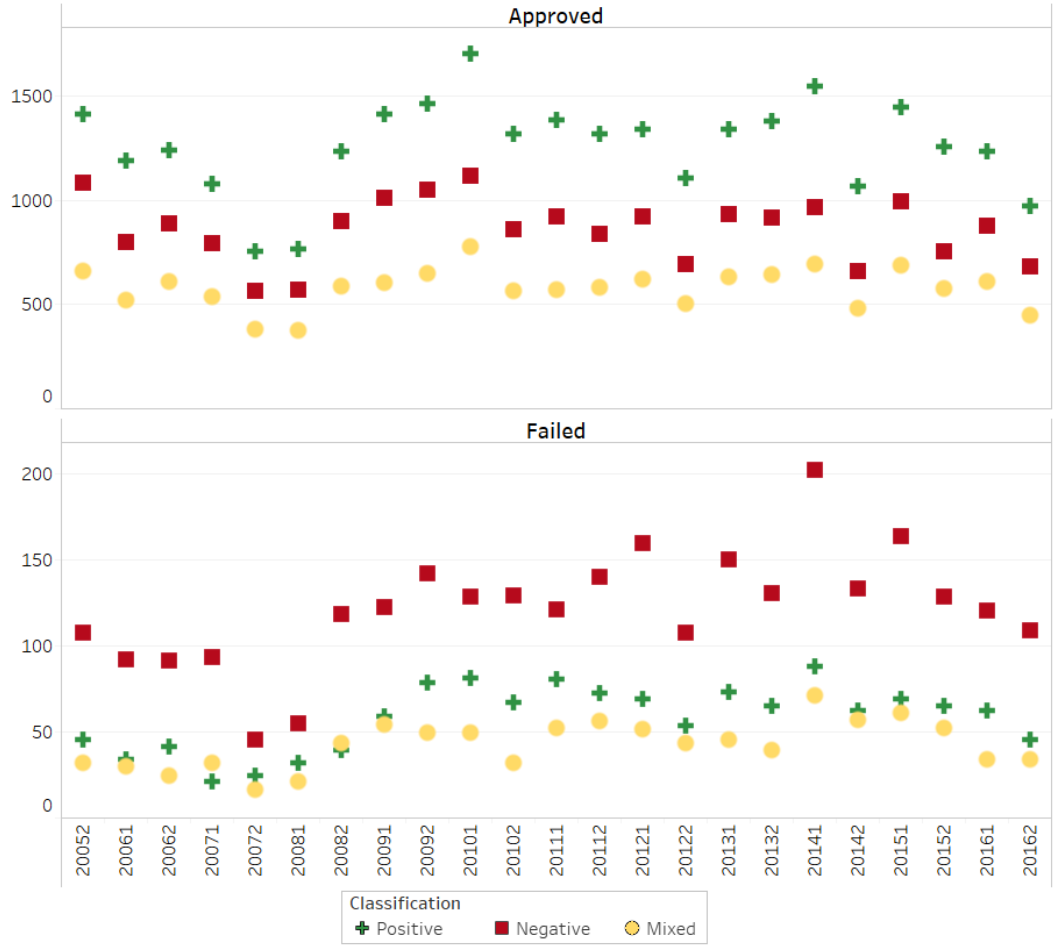

Figure 5.3: Distribution of comments by period and final status.

contrast, most of the comments of students who failed are negative. This brief analysis suggests that the type of comment made by students is correlated with the final status they achieved.

5.3 A Neural Model Approach

5.3.1 Overview of the BERT Polarity Classification Model

The BERT Polarity Classification Model encodes each comment into a 768-dimensional embedding and has a dense layer that transforms the embeddings into a three-dimensional vector, which indicates the probability that the comment belongs to each of the three classes - "positive", "negative" or "neutral". We adopted the BERT-Base (described in Section 3.4.1, Multilingual Cased version[3] (for 104 languages), which is required since our data are written in Portuguese. To significantly speed up the training and inference with our

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

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. To take advantage of the new information generated in each iteration, the BERT layer is not frozen, that is, it is updated, capturing the new information that is generated. The output layer of the model in this case will be a vector of three values indicating the probability that the comment will be classified as "positive", "negative" or "neutral", and afterwards it will only be to select the one with the highest value among the three values that will indicate the classification of the comment. Our corresponding code is available at GitHub[4].

5.3.2 Pre-Training Step

We first executed a pre-training step to adjust the BERT model to the style of students' comments through non-annotated data. 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 teacher 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 $c[q] \in \{\text{“}negative\text{”}, \text{“}neutral\text{”}, \text{“}positive\text{”}\}$ 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 5.4 shows the distribution of the average scores obtained.

The BERT model was pre-trained with the multilingual BERT checkpoint that is publicly available and trained for 10 epochs, resulting in a newly trained checkpoint, which we simply call the *pre-trained checkpoint*.

5.3.3 Training Step

After the pre-training step, we proceeded to experiment with three setups of the model, using a 5-fold cross-validation strategy, applied to a set of 800 manually classified comments. This manual classification was carried out by two people, to reduce bias; each person individually classified the 800 comments; then there was a joint discussion about those comments whose classification did not coincide; the agreement index was 89%.

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

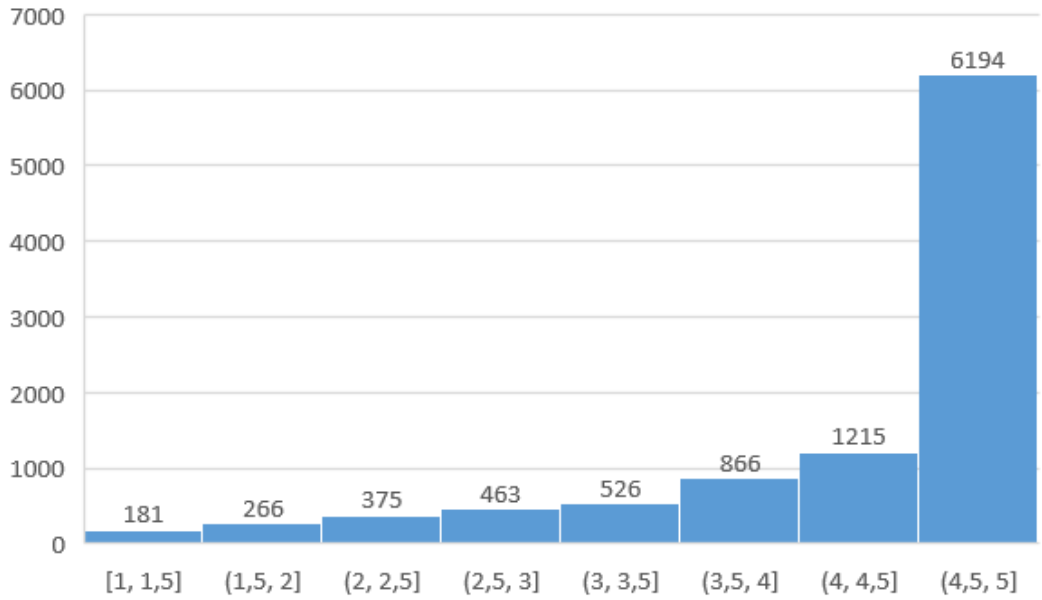

Figure 5.4: Distribution of the average score of all questions of a questionnaire from 2018.

The 800 manually annotated comments were obtained as follows. From the course surveys of the two semesters of 2019, 800 questionnaires with non-empty teacher 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 teacher; *negative*, when the comment only criticized the teacher; and *neutral* when the comment expressed no opinion or when the comment both praised and criticized the teacher. Table 5.3 shows the number of comments in each of these classes.

Table 5.3: 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

Figure 5.5 shows that the accuracy of the automatic classification is below 0.65, which justifies the need to continue with the idea of the neural model.

Therefore, each round of cross-validation used 640 comments for training and 160 comments for testing. The three setups we used were as follows:

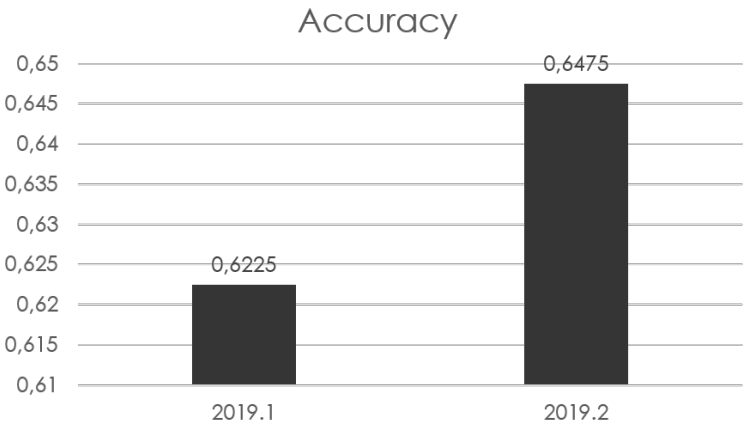

Figure 5.5: Accuracy of the comments that coincide in sentiment for the manual and automatic classification of Table 5.3

- *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 5.4 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 to obtain top-quality results.

We also adopted the Fisher-Irwin test (ROSS, 2020) to examine the hypothesis that the fine-tuned model does not have an equivalent classification performance when compared to both the zero-shot and the from-scratch models. For this purpose, we computed the Fisher-Irwin test twice. In the first

Table 5.4: Results of the setups

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**

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*). 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 these results, 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 the alternative hypotheses.

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 5.6 and 5.7 depict the results.

Figure 5.6 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 5.6 also shows that: (i) using just 40 manually annotated com-

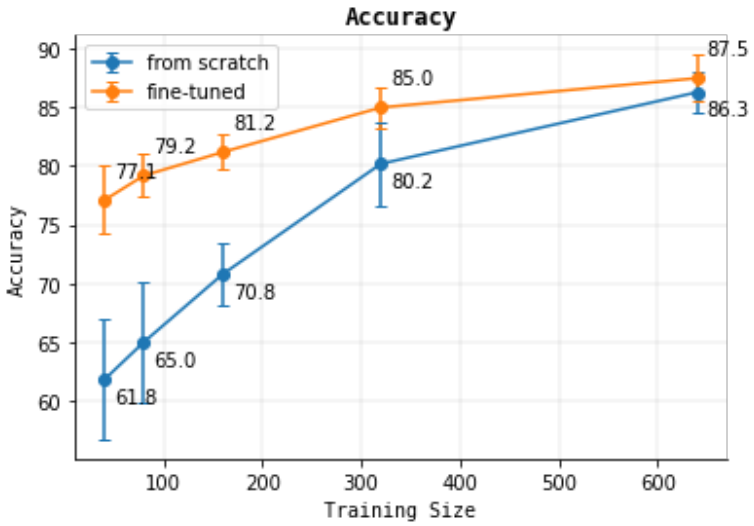

Figure 5.6: Accuracy for *From scratch* and *Fine-tuned* using train set of 40, 80, 160, 320 and 640 comments.

ments for training, the fine-tuned model achieved an average accuracy of 77.1%, while the from-scratch model 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 much better accuracy than that of the zero-shot model, shown in the first line of Table 5.4 (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.

5.3.4 Predictions

This section first applies the fine-tuned model, the best performing model, to classify the full set of comments from the 2020.1, 2020.2, and 2021.1 surveys, and the set of comments from 2019.1 and 2019.2 that were not manually classified. Then, it adds the manually classified comments from 2019.1 and 2019.2 to obtain the final distributions for all semesters, shown in Figure 5.8.

For comparison purposes, Figure 5.8 includes the distributions of the

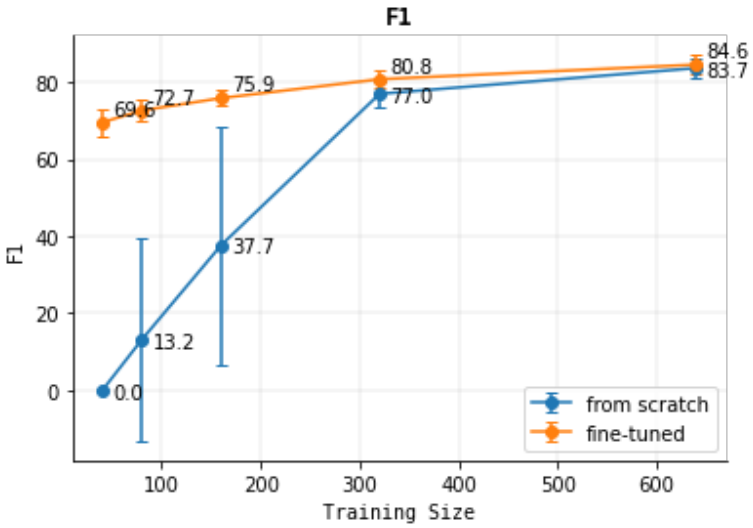

Figure 5.7: F1 for *From scratch* and *Fine-tuned* using train set of 40, 80, 160, 320 and 640 comments.

comment classifications induced by the score of *Question O: "Overall evaluation of the teacher"*, where scores 1 and 2 indicate "negative", 3 for "neutral", and 4 and 5 for "positive", considering only questionnaires with a non-empty comment about teacher performance. 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' comment sentiments and the scores of Question O indicate that students evaluated teacher performance better in 2020.1 (the early-COVID scenario) than in the other semesters. This suggests that students acknowledged the effort teachers did to keep classes running during 2020.1 and that the enthusiasm continued throughout 2020.2 (late-COVID scenario). Furthermore, students evaluated teacher performance better in 2020.1, 2020.2, and 2021.1 (online classes), by a margin of nearly 10%, when compared with 2019.1 and 2019.2 (in-person classes), respectively.

After analyzing the predictions for the set of all comments per semester from 2019, we decided to analyze in more in detail the set of comments from three major courses, one for each of the three largest centers of the PUC-Rio: Design (CTCH), Law (CCS), and Industrial Engineering (CTC). Figure 5.9 shows the distribution of the number of comments by course, period and classification. As in the previous analyzes, in the case of positive comments, the hypothesis columns are higher than the prediction columns, contrasting with the case of negative comments, where the prediction columns are higher, and the case of neutral comments, where the heights of the columns are close.

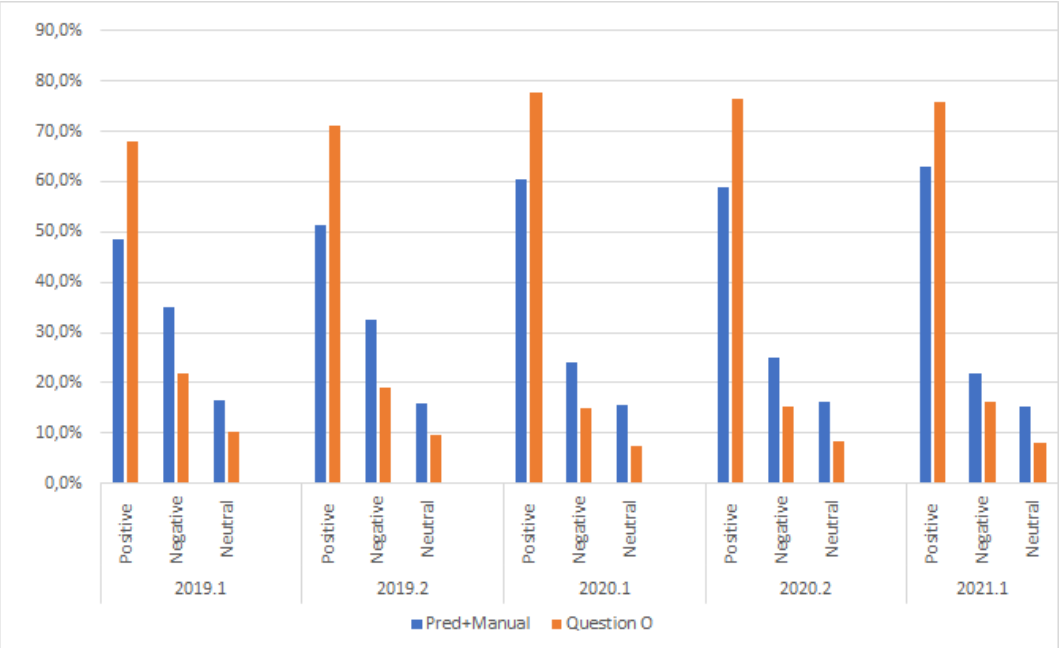

Figure 5.8: Distribution of the final classification of the comments from all surveys, using the fine-tuned model, added to the manually classified comments from 2019.1 and 2019.2 (shown in blue), and the classification of the comments from all surveys, using the score of Question O (shown in orange).

We also applied the prediction model to the comments before 2019. However, we recall that the questionnaires for these years had only one text box for comments, that is, a comment may refer to both the teachers and the course, and not just to the teacher. Figure 5.10 shows the results of such predictions. As we can observe the same distribution behavior of the comments that we observe in figures 5.8 and 5.9

5.4 Chapter summary

In this chapter, we presented two approaches to analyze the sentiment of the students' comments towards the teachers.

The first approach, described in Section 5.2, is based on a manually created dictionary that lists terms that represent the sentiment to be detected in the students' comments. This approach was implemented as a tool, called *CourseObservatory*, which classifies the polarity of a set of students' comments and helps answer a set of questions that course coordinators (or department directors) may find useful. The results were published in Jiménez et al. (2019), with data until the second semester of 2018, and indicated that the *CourseObservatory* tool outperformed a baseline tool.

This first approach, as we explained, is a fairly simple and naive solution. Furthermore, as we saw in Table 5.1, we could not classify all comments. To

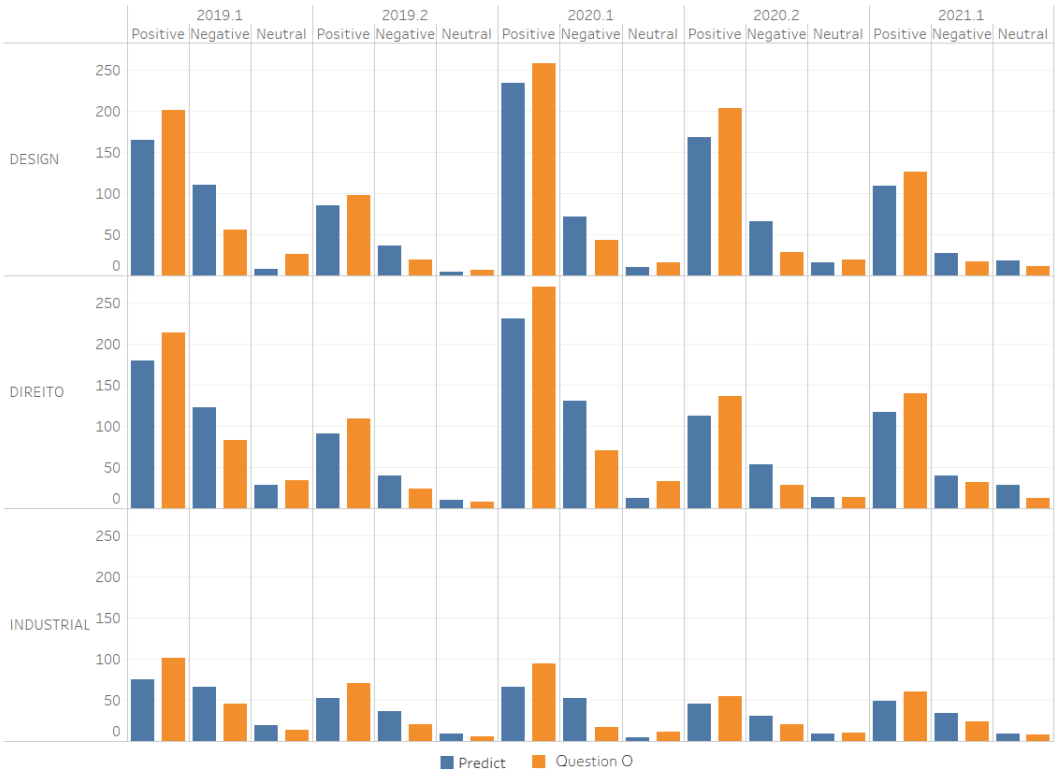

Figure 5.9: Distribution of the final classification of the comments from all surveys, using the fine-tuned model (shown in blue), and the classification of the comments from all surveys, using the score of Question O (shown in orange) for the courses Desing, Law, and Industrial Engineering.

avoid this problem, it would be necessary to analyze them and look for terms that represent sentiment and include them in the dictionary, restarting the process.

The second approach, covered in Section 5.3, was based on the BERT language representation model, and does not depend on a manually created dictionary. The model was implemented using KERAS, running on GPUs. Three types of setup were tested – zero-shot, from-scratch and fine-tuned – where the latter outperformed the other two. The fine-tuned model was then applied to predict the sentiment of the comments contained in the database since 2005.2. The results were published in Jiménez et al. (2021) and had as motivation to investigate how students reacted to the move to online classes forced by the COVID-19 pandemic, using data from 2019, 2020, and 2021. The results indicated that the second approach achieved very good performance, even when the set of manually annotated comments is small.

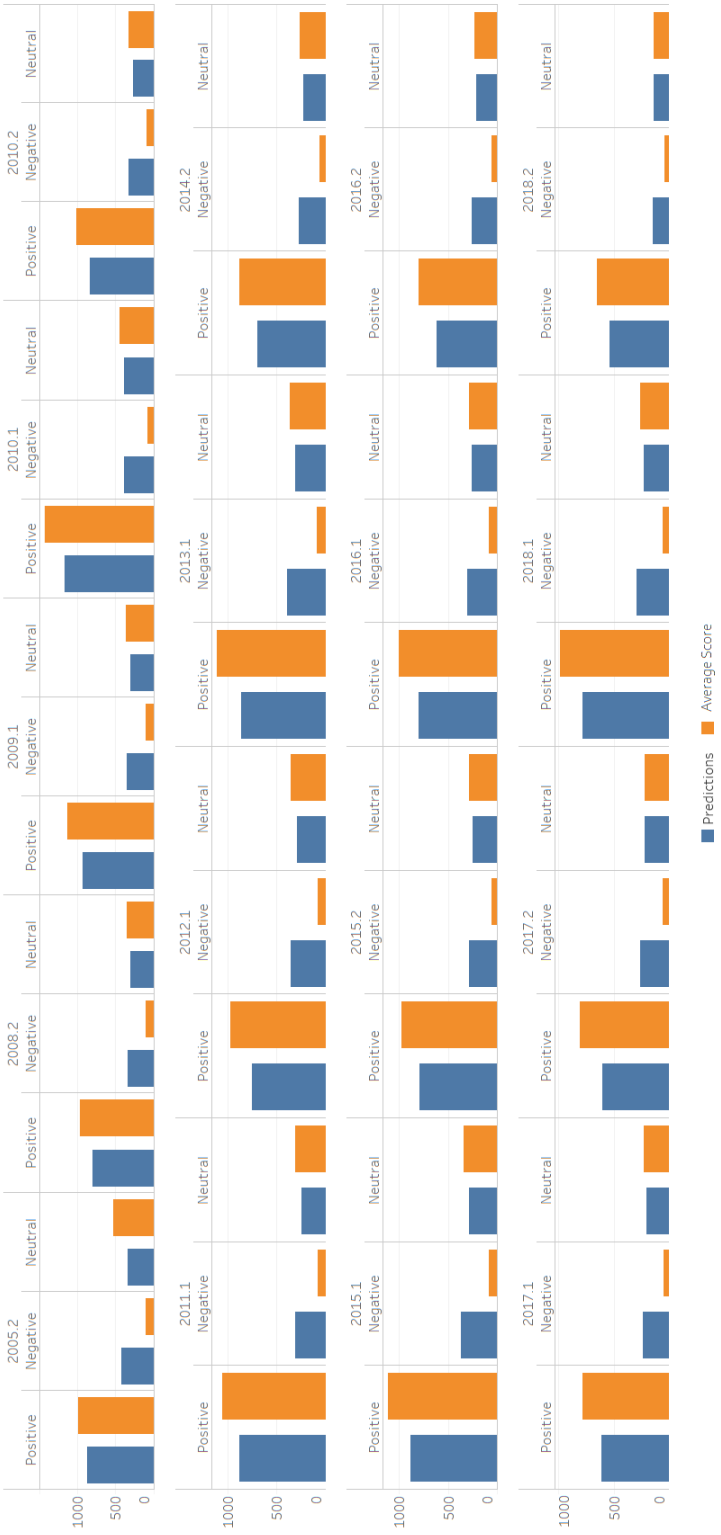

Figure 5.10: Distribution of the final classification of the comments from all surveys, using the fine-tuned model (shown in blue), and the classification of the comments from all surveys, using the average score of all questions of a questionnaire (shown in orange) for the semester until 2019 for all university.

6
Towards Comment Summarization

6.1
Introduction

This chapter addresses two related problems, intuitively described as follows:

Comment topic trending problem. Given a set C of comments, find a set of *topics* T that describes the comments in C.

Comment summarization problem. Given a set C of comments, find a subset $S \subset C$ such that S is much smaller than C, and S conveys approximately the same meaning as C.

The first problem allows one degree of freedom – the set of topics T – whereas the second problem constrains the summary S to be a set of sentences selected from C. Therefore, this chapter does not consider the problem of synthesizing a set of sentences (not in C) that concisely describe C.

This chapter investigates several strategies to address both problems in the context of students' comments about their teachers, obtained from student surveys, as described in Chapter 4. The motivation lies in the difficulty course coordinators (or department directors) have to extract useful information about the teachers of their courses (or departments) from a large set of comments. Therefore, designing a tool that summarizes a set of comments, eliminating redundant comments, would be helpful. The tool would offer course coordinators a summary of the performance of each teacher, from the students' perspective.

As reviewed in Section 2.6, different strategies and techniques have been developed to summarize the text. We could directly apply one of the text summarization strategies to a set of students' comments simply by concatenating the comments into a single text. However, one must bear in mind that the students' comments are not written by a single person, but rather they are small sentences written by different people and may repeatedly convey the same ideas. For these reasons, we discarded this approach early on.

Alternatively, we could adopt a tweet summarization technique since, just like a set of students' comments, a set of tweets is also written by several people and tweets are short sentences. This chapter explores this similarity to create strategies to address the comment topic trending problem.

This chapter is structured as follows (see Figure 6.1, discussed at the end of this section). Section 6.3 describes two strategies to address the comment topic trending problem. The other sections investigate strategies to address the comment summarization problem, that follow three basic approaches, used in isolation or in different combinations:

Partitioning Approach

Input. A set C of comments.

Partition. Partition C into (small) subsets $C_1, \ldots, C_n$.

Partial Summarization. Select a small set S_i of comments from C_i such that S_i summarizes C_i, for $i = 1, \ldots n$.

Output. Return the union S of the sets $S_1, \ldots, S_n$.

The intuition behind the **Partitioning Approach** is that the comments in a partition C_i should be redundant, that is, they should convey approximately the same meaning. Since it is difficult to argue that a set of comments is redundant, the **Partition** Step was implemented using essentially syntactical strategies. Section 6.4.1 investigates *Clustering*, that is, grouping comments by applying a clustering algorithm, based on a comment similarity measure, and Section 6.4.2 explores *Attribute Partitioning*, that is, grouping comments that have the same values for one or more attributes. In both cases, the **Partial Summarization** Step was implemented using the centroid-based summarization algorithm (CBSA for short), described in Section 3.3.2.

Ranking Approach

Input. A set C of comments, and a limit k.

Ranking. Rank C into a list of comments $c_1, \ldots, c_m$.

Output. Return the top-k comments $c_1, \ldots, c_k$.

The intuition here is that the top-ranked comments are the most important ones. The **Ranking** Step was implemented using TextRank, introduced in Section 3.2.5. The direct application of TextRank (node Top-k TextRank), discussed in Section 6.5.2, proved not to be adequate, so it was combined with clustering and the centroid-based summarization algorithm in Section 6.5.3.

Entailment Approach

Input. A set C of comments.

Entailment. Compute the *entailment graph* $G = (C, E)$ such that (c_i, c_j) is in E iff c_i entails c_j.

Simplification Compute the strongly connected components of G, and collapse each component into a single node (i.e., comment), creating a new graph $G^c = (C^c, E^c)$

Output. Return the comments $c_1, \ldots, c_k$ such that c_i does not entail any other comment in G^c (i.e., c_i is a sink of G^c).

The intuition here is that a comment c_i summarizes all comments c_j such that c_j transitively entails (i.e., implies) c_i. The **Entailment** Step was implemented using a specially trained BERT model, described in Section 6.6.1. The **Simplification** Step is just an intermediate step to eliminate circular entailments. Again, the direct application of this approach proved not to be adequate and is not described in this chapter. Section 6.6.2 discusses how to use entailment in combination with the **Ranking Approach**, clustering and the centroid-based summarization algorithm (node "Entailment + TextRank + Clustering + CBSA").

Figure 6.1 depicts the strategies to address the comment topic trending and the comment summarization problems investigated in this chapter. The second level nodes correspond to the sections of this chapter and are labeled with the approaches described above; the leaves correspond to the strategies investigated and are labeled with the names of the strategies. Briefly, this chapter experimented with the following strategies.

Comment topic trending problem

- the Market-Basket Analysis and the Topic Modeling strategies (nodes "Market Basket Analysis" and "Topic Modeling", respectively).

Comment summarization problem

- Clustering combined with the centroid-based summarization algorithm (node "Clustering + CBSA").
- Partitioning by sentiment analysis or by overall teacher score combined with the centroid-based summarization algorithm (nodes "Partitioning by Sentiment + CBSA" and "Partitioning by Score + CBSA", respectively).
- the Top-k TextRank Approach (node "Top-k TextRank").
- TextRank combined with clustering and the centroid-based summarization algorithm (node labeled with "TextRank + Clustering + CBSA").
- the Entailment Approach, combined with TextRank, clustering, and the centroid-based summarization algorithm (node "Entailment + TextRank + Clustering + CBSA").

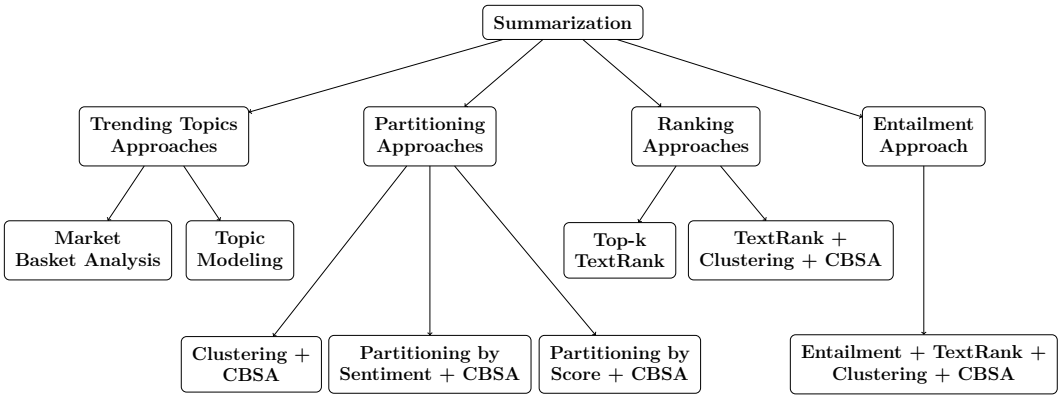

Figure 6.1: Summarization strategies.

Finally, Section 6.2 describes the set of comments used to experiment with the various strategies, while Section 6.7 presents the results of a comparison between the strategies, using the ROUGE metric, reviewed in Section 2.6.

6.2 Use of the Course Survey Data

By analyzing the course survey data, described in Chapter 4, we observed that summarizing a large set of comments was not effective, basically because it hardly makes sense to summarize comments from different departments or courses, or a group of teachers, or a long list of semesters. For example, it is not helpful to return the comment "Não é um bom professor" ("He is not a good teacher") as a summary of a set of comments for a group of 50 teachers, say, since it was not possible to identify to which specific teacher the comment was about, or to infer that the comment was about the entire set of teachers. It is also reasonable to summarize comments on a per-semester basis because teachers can vary their performance over the years, which would be hidden by a summary obtained from a list of comments spanning several semesters. The last filter to consider is to group comments by the same course and teacher, since a teacher may be responsible for several courses and his/her performance may also vary from course to course.

For the experiments with the strategies that we present in this chapter, we will then use a set of comments about the same teacher, semester, and course. We randomly selected a teacher with at least 10 comments, as shown in Table 6.1.

As each comment can represent several ideas, we separate each comment into sentences, using punctuation marks, and treat each sentence as a different comment. Table 6.2 shows the sentences obtained from the comments listed in Table 6.1.

Table 6.1: Comments selected for the experiments.

index	Comment
1	Admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom. As aulas foram ótimas.
2	Excelente professor! Atencioso, próximo ao aluno, didático.
3	Prestativo, e de uma competência e simplicidade extraordinária.
4	Excelente professor
5	Bom professor, mas não deixa o aluno repousar um pouco no intervalo.
6	Professor excelente!
7	O professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis. Excelente.
8	Muito bom, não mediu esforços para ajuadar no desmpenho dos alunos.
9	O PROFESSOR É MUITO BOM E INTELIGENTE, MAS INFELIZMENTE NÃO CONSEGUIU ACOMPANHAR AS DEMANDAS DE ATUALIZAÇÕES PARA EAD.
10	Maravilhoso!

Table 6.2: Sentences obtained from the selected comments.

New index	**Old index**	**Comment**
1	1	Admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom.
2	1	As aulas foram ótimas.
3	2	Excelente professor!
4	2	Atencioso, próximo ao aluno, didático.
5	3	Prestativo, e de uma competência e simplicidade extraordinária.
6	4	Excelente professor
7	5	Bom professor, mas não deixa o aluno repousar um pouco no intervalo.
8	6	Professor excelente!
9	7	O professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis.
10	7	Excelente.
11	8	Muito bom, não mediu esforços para ajuadar no desmpenho dos alunos.
12	9	O PROFESSOR É MUITO BOM E INTELIGENTE, MAS INFELIZMENTE NÃO CONSEGUIU ACOMPANHAR AS DEMANDAS DE ATUALIZAÇÕES PARA EAD.
13	10	Maravilhoso!

To be able to compare the summaries obtained by each strategy, using the ROUGE metric (see Section 2.6), we need a *reference summary*, constructed by a person who is supposed to be able to semantically analyze the comments. The task of constructing a reference summary is not straightforward, though, because the students may write comments from different perspectives, or may even write contradictory comments. For example, observe that in Table 6.2:

– Comments 2 and 3 express the same idea, although one focuses on the teacher's point of view, while the other talks about the classes.

– Comments 1 and 12 contradict each other – one says that the teacher

adapted to the new technologies, while the other says the opposite.

After these considerations, we asked three people to create a reference summary for the comments in Table 6.2. We asked more than one person to try to reduce subjectivity. The summaries are (the number in parenthesis is the new index of the sentence in Table 6.2; it is included to help the reader and should not be considered as part of the summary):

1. Reference Summary 1:
   (1) admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom;
   (7) bom professor, mas não deixa o aluno repousar um pouco no intervalo;
   (12) o professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.

2. Reference Summary 2:
   (1) admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom;
   (6) excelente professor;
   (11) muito bom, não mediu esforços para ajudar no desempenho dos alunos.

3. Reference Summary 3:
   (1) admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom;
   (9) o professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis;
   (12) o professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.

These reference summaries were created with sentences selected from the set of comments. Indeed, in this research, we limit ourselves to summarizing a set of comments using sentences extracted from the set (and not rewriting a summary with new sentences). Section 6.7 will use the reference summaries to compare the summaries obtained by each strategy.

6.3 Trending Topics Approaches

This section describes strategies that, although not specifically designed to summarize a set of comments, can be used to classify the comments and, together with a sentiment analysis (Section 5.3.1), produce sentences such as "They speak negatively about the teacher". These strategies were based on the Market Basket Analysis and Topic Modeling.

6.3.1 Market Basket Analysis

The idea of Market Basket Analysis is to determine how often two or more objects co-occur and, once the frequency of such sets is found, to extract association rules between the objects. It can be applied to different scenarios, beyond the supermarket scenario from which its name is derived. For more details, we refer the reader to Agrawal, Imieliński & Swami (1993), Tan, Steinbach & Kumar (2005), Han, Kamber & Pei (2012).

To propose a Market Basket Analysis strategy to summarize comments, we manually classified 1,150 comments, chosen as follows. From the course surveys of the two semesters of 2019, 800 questionnaires to evaluate teacher performance and 350 questionnaires to evaluate courses, with non-empty 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 about the teacher (5 * 5 * 16 = 400) and 7 closed-ended questions about the discipline (5 * 5 * 7 = 175) in each of the semesters ((400 + 175) * 2 = 1,150).

For each of the comments, we marked the topics that were discussed. The final topics were: *teacher, discipline/content, schedule, bibliography, tests, exercises, support tools, monitoring*, and *others*. To apply the Market Basket Analysis, we filtered the comments for each of the topics and we checked in those comments which were the most frequent terms that represent the topic. Figure 6.2 summarizes the results of these analyzes.

The final idea would then be to search for these terms in the remaining comments and thus be able to deduce what is being expressed in each comment. There may be words that are more frequent in some topics than in others. For example, "aulas", which is in the topics *teacher* and *monitoring*, but has more value than the first one; it will then be a word used to represent *teacher*; when it appears with the word "falta", it will represent *monitoring*.

Then, applying this technique to the comments of the selected teacher, we can see in Table 6.3 the classification of the topic to which the comment

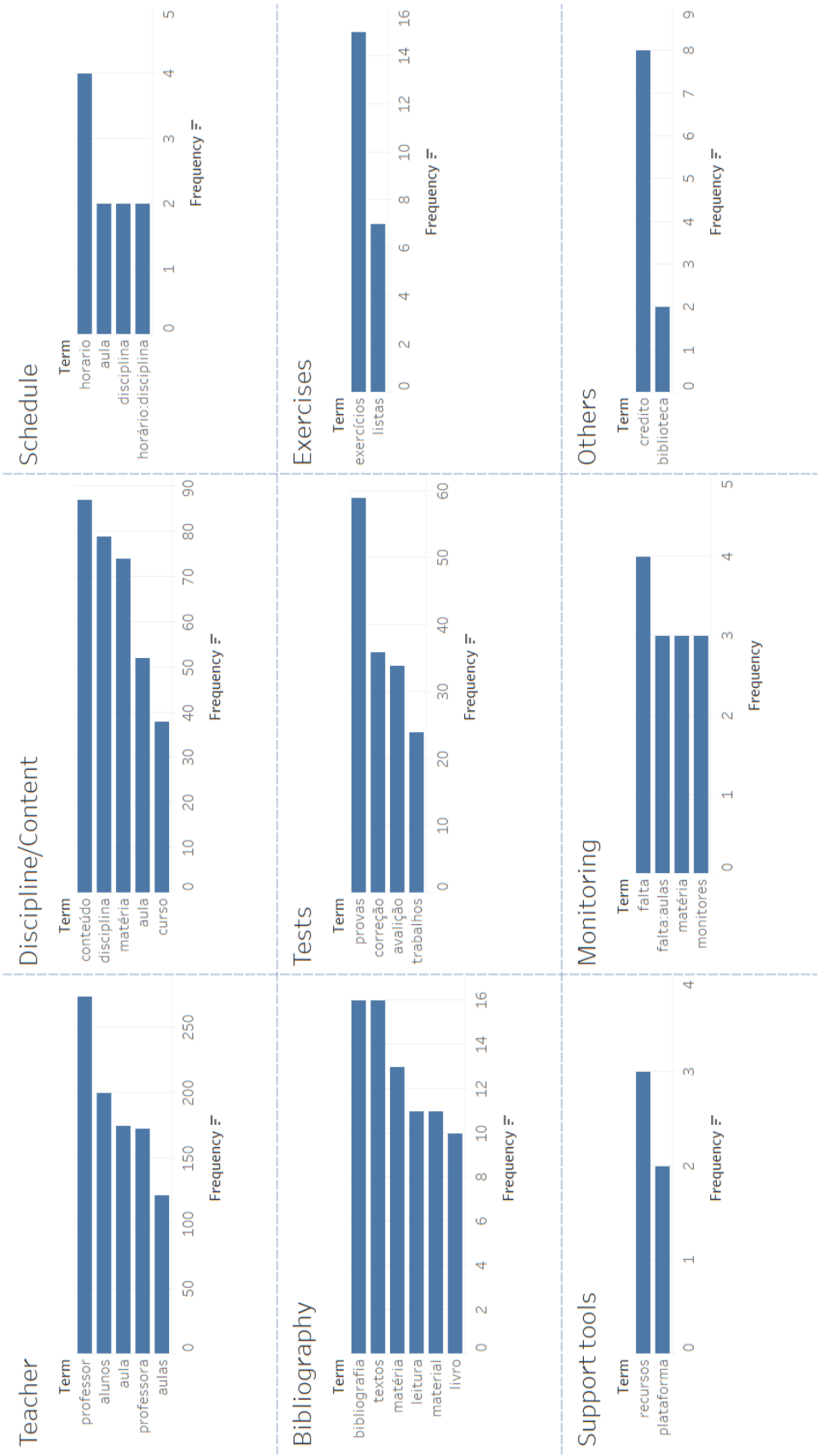

Figure 6.2: Frequent terms for a topic. Each scale is different because the number of comments analyzed is different due to the filter applied.

belongs by the words it has. As can be seen, some of the comments were not classified and this is due to the fact that the students omitted the subject to which they refer.

Table 6.3: Comment topics using Market Basket Analysis

New index	Comment	Topic
1	Admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom.	Teacher and Support tools
2	As aulas foram ótimas.	Teacher
3	Excelente professor!	Teacher
4	Atencioso, próximo ao aluno, didático.	
5	Prestativo, e de uma competência e simplicidade extraordinária.	
6	Excelente professor	Teacher
7	Bom professor, mas não deixa o aluno repousar um pouco no intervalo.	Teacher
8	Professor excelente!	Teacher
9	O professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis.	Teacher, Test, Discipline and Support tools
10	Excelente.	
11	Muito bom, não mediu esforços para ajuadar no desmpenho dos alunos.	
12	O PROFESSOR É MUITO BOM E INTELIGENTE, MAS INFELIZMENTE NÃO CONSEGUIU ACOMPANHAR AS DEMANDAS DE ATUALIZAÇÕES PARA EAD.	Teacher
13	Maravilhoso!	

This strategy, although effective, is limited to the topics that were manually classified. However, such topics may become obsolete, as something new begins to be commented by the students and therefore must be taken into account.

6.3.2 Topic Modeling

Topic models are a type of statistical language models that are used to discover hidden structures in a collection of texts; these topics should be those that best represent the information contained in the collection of texts. With this definition, we considered applying topic models to our dataset, thus obtaining the topics that would represent each teacher. As in the other experiments in this chapter, we will filter and work with one teacher at a time and we will also clean the stop words and lower-case the text.

One of the characteristics of the topic modeling algorithm is to define the number of topics to be created. For the teacher that we experiment with we selected 5 topics, the results can be seen in Figure 6.3

Each topic is defined by a combination of keywords, and each keyword contributes with a certain weight to the topic. From the keywords, we might induce what is being talked about and define a topic. Figure 6.3 illustrates how the topic modeling algorithm was applied to the comments of

Figure 6.3: Keyword and weight with which it contributes to the topic.

a single teacher. In this case, we should interpret the topics as if they were characteristics of the teacher. We therefore ignore the nouns and verbs, and focus on the adjectives. which leads to the following topics to model this teacher:

0 - Excelente
1 - Solicito, disposto
2 - Esforços
3 - Maravilhoso
4 - Bom

This solution proved useful to define the keywords that represent a teacher, but it loses the context of the keywords, which can be relevant information.

6.3.3 Lessons Learned from the Trending Topics Approaches

The Market Basket Analysis and Topic Modeling approaches do not summarize a set of comments, but they allow us to classify comments into topics.

To apply the Market Basket Analysis approach, we need to separate and classify a set of comments by topics. This is necessary to uncover the most frequent words which are then used to classify the rest of the comments. However, students often omit the subject they are talking about so that some comments remain unclassified.

Although the Topic Modeling approach is useful to define the topics in a given scenario, it does not capture important information because it may be necessary to eliminate words that are often repeated, such as "professor" or "professora", which implies that the context can be lost.

6.4 Partitioning Approaches

This section investigates strategies to summarize a set of comments based on the idea of clustering the comments, using a similarity measure, or by partitioning comments by the values of one or more attributes. The centroid-based summarization algorithm is used to summarize the comments in each cluster or group.

6.4.1 Clustering combined with the Centroid-based Summarization Algorithm

To compute the similarity between pairs of comments, we proceed in three steps. First, we use BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) as the sentence transformer model. BERTimbau is a repository that contains pre-trained BERT models for Portuguese. Specifically, we used *'neuralmind/bert-base-portuguese-cased'*, a variant trained on the BrWaC (WAGNER et al., 2018) corpus, a large Portuguese corpus, for 1,000,000 steps, using a whole-word mask.

Then, we use the *encode* function provided by SBERT (see Section 3.4.2). The encode function creates an array with the embeddings corresponding to each comment in the dataset, as in the following example:

Sentence: “Solícito, didático, pontual e sempre bem disposto e bem humorado.”
Embedding: [4.54094727e-03 -1.58600658e-01 3.08290869e-01 3.43252301e-01 4.65752035e-01 1.41963080e-01 -2.53006786e-01 -1.36909392e-02 2.75858790e-01 -4.11321312e-01 1.81449503e-01 6.69072032e-01 5.52341230e-02 -1.50158510e-01 -3.02290440e-01 -2.90237725e-01.....]

The third step is to compute the similarity between pairs of encoded comments using the cosine similarity function recommended by SBERT and described in Section 3.2.3.

With the similarity computed for all the pairs of comments, we group similar comments using a clustering algorithm, such as KMeans (see Section 3.3.3).

Finally, we summarize the set of comments in each cluster, using the centroid-based summarization algorithm, and combine all cluster summaries into a final summary.

The experiments indicated that this strategy suffers from two related problems:

– comments with high similarity might be semantically very different; for example, the comments "Não é um bom professor" ("He is not a good teacher") and "É um bom professor" ("He is a good teacher") were 0.97% similar.

– to avoid this problem, we had to tight the clustering criteria, thereby creating too many small clusters, which in turn led to too many comments in the final summary.

A better strategy would be to directly partition the set of comments, using attribute values, as discussed in the next section.

6.4.2 Attribute Partitioning

This section describes experiments with the direct application of the centroid-based summarization algorithm, described in Section 3.3.2, to (small) sets of comments obtained by different partitioning schemes.

As remarked in Section 6.2, summarizing a large, heterogeneous set of comments is not reasonable. The set should first be partitioned by teacher, discipline, and semester. Another helpful partitioning schema would be by the polarity of the comments – positive, negative, and neutral – taking advantage of the result obtained in Section 5.3.1. A third possibility would be to partition a set of comments by the overall teacher score so that comments in each partition hopefully reflect a common opinion of the students about the teacher.

We then partition comments by the following criteria:

– *teacher/discipline/semester*, when the comments are partitioned by teacher, discipline, and semester

– *predicted sentiment polarity*, when the comments are partitioned by their predicted sentiment polarity, as Positive, Negative, and Neutral

– *overall teacher score*, when the comments are partitioned by the overall score the student gave to the teacher, which is obtained from the questionnaire

Table 6.4 shows the sentiment prediction and the overall teacher score for each comment listed in Table 6.2. Column Prediction represents the sentiment polarities “Positive”, “Negative”, and “Neutral” as 2, 0, and 1, respectively.

Table 6.4: Comments of the teacher for the experiments.

index	Comment	Overall teacher score	Predicted sentiment polarity
1	Admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom.	4.0	2
2	As aulas foram ótimas.	4.0	2
3	Excelente professor!	5.0	2
4	Atencioso, próximo ao aluno, didático	5.0	2
5	Prestativo, e de uma competência e simplicidade extraordinária.	5.0	2
6	Excelente professor	5.0	2
7	Bom professor, mas não deixa o aluno repousar um pouco no intervalo	2.0	0
8	Professor excelente!	5.0	2
9	O professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis.	5.0	2
10	Excelente.	5.0	2
11	Muito bom, não medi esforços para ajuadar no desmpenho dos alunos.	5.0	2
12	O PROFESSOR É MUITO BOM E INTELIGENTE, MAS INFELIZMENTE NÃO CONSEGUIU ACOMPANHAR AS DEMANDAS DE ATUALIZAÇÕES PARA EAD.	1.0	1
13	Maravilhoso!	5.0	2

6.4.2.1 Partitioning by Sentiment combined with the Centroid-based Summarization Algorithm

In the first experiment, we partitioned the comments in Table 6.4 by the sentiment polarity and separately applied the centroid-based summarization algorithm to each group, if the partition had at least one comment, to obtain a text summary for the group. The text summary for the full set of comments is the concatenation of the text summaries obtained for each group.

The text summaries for each group of comments (i.e., for the comments with the same sentiment polarity) were:

0
Bom professor, mas não deixa o aluno repousar um pouco no intervalo
1
O professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.
2
Excelente professor

Then, the text summary for the full set of comments simply is (again, the number in parenthesis is the new index of the sentence in Table 6.2; it is included to help the reader and should not be considered as part of the summary):

– Partitioning by Sentiment Summary:
(7) Bom professor, mas não deixa o aluno repousar um pouco no intervalo;
(12) O professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead;
(6) Excelente professor.

6.4.2.2 Partitioning by Overall Teacher Score combined with the Centroid-based Summarization Algorithm

In the second experiment, we grouped the comments in Table 6.4 by the overall score the student gave to the teacher, obtained from the questionnaire, and again separately applied the centroid-based summarization algorithm to each group, if the group had at least one comment, to obtain a text summary for the group. The text summary for the full set of comments is the concatenation of the text summaries obtained for each group.

The text summaries for each group of comments (i.e., for the comments with the same overall teacher score) were:

1
O professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.
2
Bom professor, mas não deixa o aluno repousar um pouco no intervalo
3
(No comments for this classification)
4
As aulas foram ótimas.
5
Excelente professor

Then, the text summary for the full set of comments is:

– Partitioning by Overall Teacher Score:
(12) O professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead;
(7) Bom professor, mas não deixa o aluno repousar um pouco no intervalo;
(2) As aulas foram ótimas;
(6) Excelente professor.

6.4.3 Lessons Learned from the Partitioning Approaches

The experiments indicated that the clustering strategy had two limitations. First, the similarity between the two sentences proved not to be adequate. For example, the sentences "Não é um bom professor" ("He is not a good teacher") and "É um bom professor" ("He is a good teacher") were 0.97% similar, which forced the clustering algorithm to assign them to the same cluster, which is not reasonable. Second, the clusters were fairly small and numerous, so that the final summary was not adequate.

The experiments with attribute partitioning and the direct application of the centroid-based summarization algorithm proved adequate to summarize (small) sets of comments, obtained by different partitioning schemes, easily computed from the questionnaires or by the predicted sentiment polarity of the comments. This strategy required the manual definition of the correct partitioning schema, though. Furthermore, the number of sentences in the summary at most the number of different values of the attributes used for partitioning.

6.5 Ranking Approach

6.5.1 The TextRank Algorithm Revisited

As described in Section 3.2.5, TextRank is a text ranking technique, which builds upon the idea of PageRank. In outline, the TextRank algorithm is as follows:

1. Process the text eliminating stop words, minimizing the whole sentence, eliminating special characters.

2. Find a vector representation for each sentence.

3. Compute the similarity between the sentences and store the values in a square similarity matrix.

4. Apply the PageRank algorithm as usual but using the similarity matrix.

Step 1 is quite simple since it only requires applying string functions, as defined for example in the `pandas` (MCKINNEY, 2010) Python library.

Step 2 requires creating a vector that represents each sentence. We used `Word2Vec` (ŘEHŮŘEK; SOJKA, 2010) Python library, which returns an array where each sentence is represented by a word embedding.

Step 3 requires computing the similarity between the vectors that represent the sentences. We adopted the cosine similarity, explained in Section 3.2.3. The result of Step 3 is a matrix with the respective cosine similarity scores for all possible sentence pairs.

Step 4 requires applying the PageRank algorithm to the matrix returned by Step 3. We implemented Step 4 using the `pagerank_numpy` function from Python library `networkx` [1], which applies the PageRank algorithm.

The code of this algorithm is in GitHub[2].

6.5.2 Top-k TextRank

The Top-k TextRank approach directly applies the TextRank algorithm to a set of comments and selects the top-k comments as the set summary.

For the experiment, we again used the comments shown in Table 6.2. The results of the first step of the algorithm are as follows:

[1] Avaible in https://networkx.org/

[2] Available at https://github.com/hguillot/textrank_algorithm

'admiro esforço humildade professor adaptar atual recurso tecnológico aulas aplicativo zoom',
'aulas ótimas',
'excelente professor',
'atencioso próximo aluno didático',
'prestativo competência simplicidade extraordinária',
'excelente professor',
'bom professor deixa aluno repousar pouco intervalo',
'professor excelente',
'professor mostrou sempre disposto ajudar solicito tirar dúvidas preparou avaliações coerentes conteúdo exposto sempre interagiu plataformas disponíveis',
'excelente',
'bom medi esforços ajuadar desmpenho alunos',
'professor bom inteligente infelizmente conseguiu acompanhar demandas atualizações ead',
'maravilhoso'

Note that the sentences are in lowercase letters, and the punctuation marks and stop words are missing.

After applying the rest of the steps of the TextRank algorithm, each sentence has a score that represents its importance, as shown in the rightmost column of Table 6.5.

Table 6.5: TextRank Experiment Results

index	Sentences	TextRank Score
8	Professor excelente!	0.124176
6	Excelente professor	0.124176
3	Excelente professor!	0.124176
1	Admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom.	0.123321
12	O PROFESSOR É MUITO BOM E INTELIGENTE, MAS INFELIZMENTE NÃO CONSEGUIU ACOMPANHAR AS DEMANDAS DE ATUALIZAÇÕES PARA EAD.	0.118808
7	Bom professor, mas não deixa o aluno repousar um pouco no intervalo	0.110866
9	O professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis.	0.087926
10	Excelente.	0.064068
11	Muito bom, não medi esforços para ajuadar no desmpenho dos alunos.	0.054296
2	As aulas foram ótimas.	0.025121
5	Prestativo, e de uma competência e simplicidade extraordinária.	0.014354
13	Maravilhoso!	0.014354
4	Atencioso, próximo ao aluno, didático	0.014354

Suppose that we select only the 3 comments with the highest scores. The summary would be as follows:

– Top-k TextRank Summary:
(8) Professor excelente!;
(6) Excelente professor;
(3) Excelente professor!.

which conveys redundant information. On the other hand, if we chose only one sentence as the summary of the set of comments to avoid this problem, we could be leaving out important information, such as some of the negative characteristics of the teacher, in this case. We then decided to apply a clustering approach to these results, as explained in the next section.

6.5.3 TextRank combined with Clustering and the Centroid-based Summarization Algorithm

To circumvent the problems with the top-k TextRank approach, we defined a second approach based on TextRank to summarize a set of comments C:

1. Compute the TextRank scores of the comments in C.
2. Cluster the comments using as comment similarity the difference between their TextRank scores.
3. Summarize each cluster, using the centroid-based summarization algorithm.
4. The summary for C would be the concatenation of the selected comments.

For the clustering technique, we used the KMeans algorithm, explained in Section 3.3.3. To find the number of clusters, we adopted the elbow strategy, as also explained in that section.

Using the same set of comments as in Section 6.5.2, Figure 6.4 shows that the appropriate number of clusters would be 2 or 3, since this is where the "elbow" of the line lies.

Choosing k = 3, the comments would be separated into 3 different clusters, as shown in Table 6.6. By applying the centroid-based summarization algorithm to the comments in each cluster, and concatenating the selected comments, the summary would be as follows:

– TextRank Clustering Summary:
(8) Professor excelente!;
(11) Muito bom, não mediu esforços para ajudar no desempenho dos

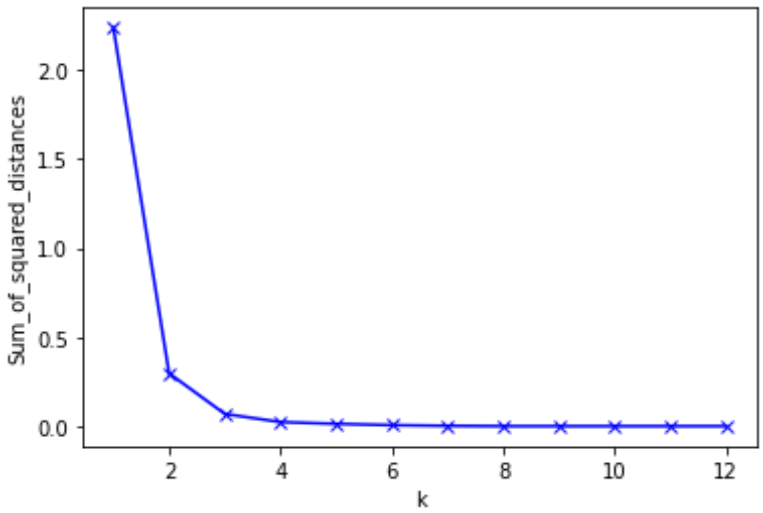

Figure 6.4: Optimal k for the data obtaining with the Elbow Method.

alunos;
(4) Atencioso, próximo ao aluno, didático.

Table 6.6: TextRank Clusters and Centroid sentences

index	Sentences	TextRank Score	Cluster
8	**Professor excelente!**	0.124176	1
6	Excelente professor	0.124176	1
3	Excelente professor!	0.124176	1
1	Admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom.	0.123321	1
12	O PROFESSOR É MUITO BOM E INTELIGENTE, MAS INFELIZMENTE NÃO CONSEGUIU ACOMPANHAR AS DEMANDAS DE ATUALIZAÇÕES PARA EAD.	0.118808	1
7	Bom professor, mas não deixa o aluno repousar um pouco no intervalo	0.110866	1
9	O professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis.	0.087926	2
10	Excelente.	0.064068	2
11	**Muito bom, não medi esforços para ajuadar no desmpenho dos alunos.**	0.054296	2
2	As aulas foram ótimas.	0.025121	0
5	Prestativo, e de uma competência e simplicidade extraordinária.	0.014354	0
13	Maravilhoso!	0.014354	0
4	**Atencioso, próximo ao aluno, didático**	0.014354	0

6.5.4 Lessons Learned from the Ranking Approaches

The top-k TextRank approach recommends selecting those comments with the highest scores as a summary of the set of comments, but it might therefore return redundant information. The refined approach first clusters the set of comments by TextRank and then summarizes each cluster by applying

the centroid-based summarization algorithm. The summary for the full set of comments is the concatenation of the selected comments.

6.6 Entailment Approach

This section explores entailment, introduced in Section 3.2.1, to define a strategy to summarize students' comments.

Our first approach was to use entailment to group comments: all comments that entail a *topic* S are assigned to the same group, and S is taken as a summary of the group. To select the set of topics, we experimented with two strategies:

- Manually define the set of topics as those that are usually mentioned in the comments in each semester;
- Choose some comments from the set of comments.

This approach proved not to be adequate: many comments were lost, so we ended up with a large number of groups, which in turn resulted in inadequately long summaries.

Our second approach was to compute the entailment for all pairs of comments, generating an entailment score matrix, which was then used to apply the TextRank strategy from Section 6.5. The following subsections detail this approach.

6.6.1 Computing entailment with BERT

We adopted BERT to compute the entailment between pairs of comments. First, we trained BERT with examples of entailment, using the ASSIN2 (REAL; FONSECA; OLIVEIRA, 2020) dataset, which classifies a pair of sentences into "entailment" or "none", we use the same structure of the model described in 5.3.1. The ASSIN2 training dataset has 6,500 rows, classified into 3,250 that have entailment and 3,250 with none.

After training the model, the next step was to predict whether a pair of sentences has entailment. Note that, given two sentences A and B, we must test if A entails B, as well as if B entails A since, as explained in Section 3.2.1, entailment is not symmetrical. The result was represented as a matrix with two columns and n lines, where n is the number of all combinations of comments. For the pair of sentences, (A, B) corresponding to line i, the first column of the line i indicates whether A entails B or not, and the second column of line i contains a score that represents the probability that A entails B.

Looking at the first column, we could create the sets with sentences that have entailment between them. But a better option was to use the second column and apply the TextRank algorithm from Section 6.5.

6.6.2 Entailment combined with TextRank, Clustering, and the Centroid-based Summarization Algorithm

We again used the comments shown in Table 6.2. Specifically in this experiment, we added the sentence "opinião sobre o professor: " to the beginning of each comment. This modification added context so that, for example, the comment "Maravilhoso" became meaningful to the entailment process.

We created all pairs of comments and computed their entailment, using the BERT model trained for entailment. The model recognized 38 entailment relationships, some of which are shown below (Appendix D contains the full list). For more clarity, we removed the added text string.

```
'atencioso, próximo ao aluno, didático. - prestativo, e de uma competência e
simplicidade extraordinária.',
'excelente professor! - professor excelente!',
'excelente professor! - maravilhoso!',
'as aulas foram ótimas. - atencioso, próximo ao aluno, didático.',
```

We applied the same strategy as in Section 6.5 to summarize the set of comments from this point on. We built a directed graph, using the comments as vertices, and considered the weight of the edges as the value of entailment probability returned by BERT. We then computed the TextRank for each comment, calculated the elbow value to know how many clusters to create, created the clusters with KMeans, and finally selected the centroids to create the summaries. Table 6.7 shows the TextRank values, the cluster each sentence belongs to, and the sentences that summarize each cluster in boldface, computed with the Centroid-Based algorithm.

The final summary is:

– Entailment TextRank Clustering Summary:
(3) Excelente professor!;
(5) Prestativo, e de uma competência e simplicidade extraordinária;
(11) Muito bom, não medi esforços para ajuadar no desempenho dos alunos.

Table 6.7: Results for Entailment Strategy

index	Sentences	TextRank Score	Cluster
4	Atencioso, próximo ao aluno, didático.	0.157763	1
3	**Excelente professor!**	0.120307	1
8	Professor excelente!	0.113457	1
6	Excelente professor	0.107534	1
10	Excelente.	0.104200	1
9	O professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis.	0.103195	1
13	Maravilhoso!	0.077679	0
5	**Prestativo, e de uma competência e simplicidade extraordinária.**	0.075664	0
2	As aulas foram ótimas.	0.063196	0
1	Admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom.	0.029191	2
7	Bom professor, mas não deixa o aluno repousar um pouco no intervalo	0.023219	2
12	O professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.	0.012595	2
11	**Muito bom, não medi esforços para ajuadar no desmpenho dos alunos.**	0.011994	2

6.6.3 Lessons Learned with the Entailment Approach

We must face three difficulties to compute entailments between the comments in a set. First, we have to find datasets prepared to train the language model. Second, since entailment is not symmetric, we have to compute entailment for all pairs of comments, which is costly. Third, we have to consider the strongly connected components of the entailment graph to avoid circular entailments.

Despite these difficulties, combining entailment with TextRank and clustering proved to be a reasonable strategy.

6.7 Evaluation of the Comment Summarization Strategies

This section compares the comment summarization strategies introduced in this chapter, using the ROUGE metric, explained in Section 3.2.4, and the three reference summaries defined in Section 6.2. Table 6.8 presents the results obtained.

Figure 6.5 shows the ROUGE metrics for each pair *<strategy summary, reference summary>*. We observe that:

– The reference summaries 1 and 3 contain the longest comments, so the comparison benefits the larger computed summaries, such as those computed by the partition approaches. The results for summary computed

Table 6.8: Reference and Strategy Summaries.

Strategy	Summary
Reference Summary 1	admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom. bom professor, mas não deixa o aluno repousar um pouco no intervalo. o professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.
Reference Summary 2	admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom. excelente professor. Muito bom, não medi esforços para ajuadar no desmpenho dos alunos.
Reference Summary 3	admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom. o professor se mostrou sempre disposto a ajudar, solicito para tirar dúvidas, preparou avaliações coerentes com o conteúdo exposto e sempre interagiu através das plataformas disponíveis o professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.
Partitioning by Sentiment + CBSA	excelente professor. bom professor, mas não deixa o aluno repousar um pouco no intervalo. o professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.
Partitioning by Score + CBSA	excelente professor. as aulas foram ótimas. bom professor, mas não deixa o aluno repousar um pouco no intervalo. o professor é muito bom e inteligente, mas infelizmente não conseguiu acompanhar as demandas de atualizações para ead.
Top-k TextRank	professor excelente! excelente professor. excelente professor!
TextRank + Clustering + CBSA	muito bom, não medi esforços para ajuadar no desmpenho dos alunos. professor excelente! atencioso, próximo ao aluno, didático
Entailment + TextRank + Clustering + CBSA	muito bom, não medi esforços para ajuadar no desmpenho dos alunos. excelente professor! prestativo, e de uma competência e simplicidade extraordinária.

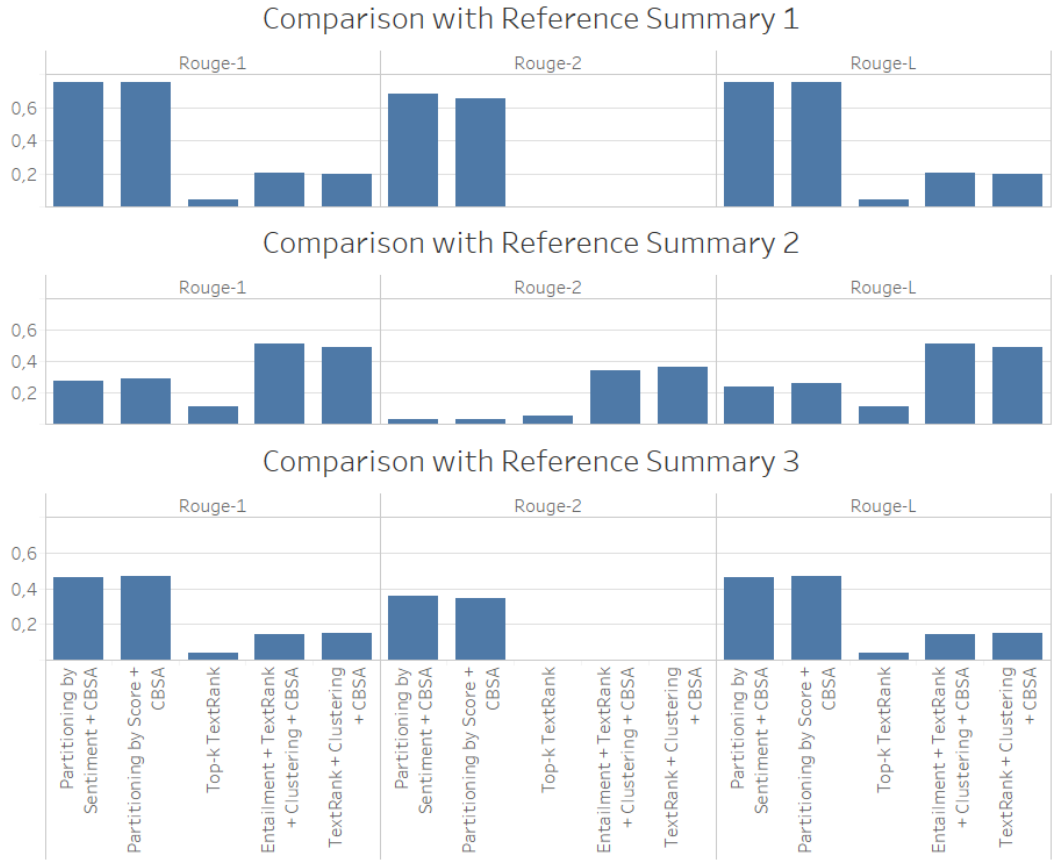

Figure 6.5: F1 measures for ROUGE-1, ROUGE-2, and ROUGE-L between the strategy summaries and the manual reference summaries.

by the top-k strategy were very low because all three comments ended up conveying the same idea. On the other hand, the results for summaries computed by the TextRank and Entailment strategies are very similar, so it cannot be concluded which of the two strategies is better.

– The reference summary 2 was a little more objective, so the best results in this comparison were obtained by the summaries computed by the TextRank and Entailment strategies, the latter being the best by 0.02 in two metrics. In the case of the partitions and top-k strategies, this time they were below the others because many of the words in the summaries computed by these strategies were not in the reference summary.

We can conclude that ROUGE, despite being the most used metric to evaluate a computed summary, does not give a decisive answer, since it will always depend on the reference summary with which the computed summary is being compared. So we can say that, except for the summary computed by the top-k strategy, the rest of the strategies compute a reasonable summary, in the sense that the computed summary represents the rest of the comments.

Given that the ROUGE measures were not conclusive, we have to take into account other characteristics to decide which is the recommended strategy. In particular, we observe that:

– the partitioning approaches require the user to have a thorough understanding of the dataset to select the appropriate filters (or attributes), and compute summaries whose sizes are limited by the number of distinct attribute values;

– the ranking approaches require computing word embeddings, the cosine similarity between the embeddings, and TextRank, which are not too costly;

– the entailment approach requires a database in Portuguese trained for BERT to find the entailment between the comments and has a high execution cost since entailment has to be computed for all pairs of comments.

Based on these observations and the results of the ROUGE metrics, we consider that the recommended comment summarization strategy is TextRank combined with clustering and the centroid-based summarization algorithm, as depicted in the gray leaf of the graph in Figure 6.6.

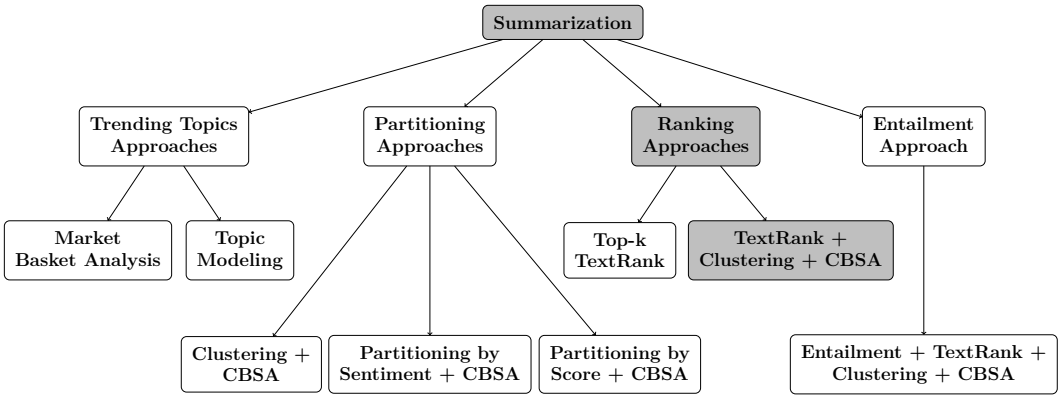

Figure 6.6: Recommended comment summarization strategy.

6.8 Further Experiments

Because of the discussion of the previous section, we apply TextRank combined with clustering and the centroid-based summarization algorithm to several sets of comments about different teachers to further assess this strategy.

– Teacher #1

Comments:

'professora extremamente prestativa, disponível para dúvidas e esclarecimentos, além disso, ela é muito inteligente , com aulas muito boas e didáticas.'
'professora maravilhosa, paciente e dedicada.'
'a professora foi impecável tanto na clareza dos conteúdos, quanto na adaptação ao ambiente online.'
'nada a declarar porque ela é perfeita.'
'desempenho excepcional.'
'as aulas de zoom tambem foram ótimas, tão boas quanto a aula presencial.'
'incrível!!'
'boa professora.'
'excelente professora!'
'ela fez muito bem o feedback das avaliações, etapa fundamental para a aprendizagem.'
'otima adaptação da matéria para plataforma ead.'
'melhor professora que tive até agora, despertou meu interesse em uma matéria que não é uma das melhores para mim mesmo em ambiente online.'

Summary:

'A professora foi impecável tanto na clareza dos conteúdos, quanto na adaptação ao ambiente online.'
'As aulas de zoom tambem foram ótimas, tão boas quanto a aula presencial.'
'Melhor professora que tive até agora, despertou meu interesse em uma matéria que não é uma das melhores para mim mesmo em ambiente online.'

– Teacher #2

Comments:

'a professora tem minha admiração, pois diante de todas as dificuldades e adversidades ela manteve a turma produzindo e até quem chegou depois, do grupo está formado, como foi o meu caso teve chace de participar de todo o processo, de aprendizagem da disciplina.'
'a professora foi excelente, exceto quando as aulas no moodle passaram a ser muito repetitivas.'
'uma ótima pessoa , compreensível e dedicada'
'apesar do contexto que estamos vivendo, conseguiu passar todo o conteúdo de maneira didática e teve uma excelente capacidade de se relacionar com os alunos, buscando novas ideias, nos ouvindo e sempre disposta a tirar dúvidas.'
'dentro da atual conjuntura, atuou com maestria com os alunos.'
'ficou monótono demais com o passar das aulas e sempre tínhamos o mesmo formato.os textos trazidos foram incríveis!'
'maravilhosa, alegre, era uma aula prazerosa.'
'a professora foi uma das professoras que mais demostraram afeiçaõe compreensão aos alunos, sempre conversando e sendo flexível para que todos os alunos conseguissem entregar as atividade, a aula no app zoom foi mantida , permitindo essa interação tão necessária entre aluno e professor, sempre abriu espaço de fala para dizer como foi a semana momento muito que do necessário para o entrosamento da turma , e toda semana havia penas atividades que resumia a matéria e assim permitia um melhor entendimento.'
'deu um suporte maravilhoso em meio a tudo isso que estamos vivendo.'
'ela despertava em nós a vontade de aprender mais, de se aprofundar no assunto e expor nossa opinião para haver a troca de interação.'
'professora maravilhosa.'
'professora maravilhosa'
'a professora foi muito atenciosa, dedicada.'
'conhece bem todos os alunos, está atenta as necessidades e desafios individuais.'
'muito bons.'
'a professora utilizou uma metodologia muito boa aplicada ao contexto que nos encontramos e motivou em nós o interesse pela disciplina.'

Summary:

'Ficou monótono demais com o passar das aulas e sempre tínhamos o mesmo formato. Os textos trazidos foram incríveis!'
'Ela despertava em nós a vontade de aprender mais, de se aprofundar no assunto e expor nossa opinião para haver a troca de interação.'
'Professora maravilhosa.'
'A professora utilizou uma metodologia muito boa aplicada ao contexto que nos encontramos e motivou em nós o interesse pela disciplina.'

6.9 Chapter summary

In this chapter, we defined and experimented with several strategies to summarize a set of comments. We also tried to validate the results of each strategy using ROUGE, the most popular metric in the summarization area, but this was not conclusive. Therefore, we had to take into account other characteristics to decide which is the recommended strategy.

Briefly, the results were:

- The Market Basket Analysis and Topic Modeling strategies do not summarize a set of comments, but they allow us to classify comments into topics. Combining this strategy with sentiment analysis, discussed in the previous chapter, we may create summaries that provide information such as "Negative/Positive Topic".
- The partitioning strategies are adequate for small sets of comments. Applying sentence similarity combined with clustering proved not to be a good strategy. Partitioning a set of comments by attribute value is a better strategy, although it requires the manual selection of the correct attributes to partition the set.
- The entailment strategy, despite being efficient when combined with TextRank, clustering, and the centroid-based summarization algorithm requires a database in Portuguese trained for BERT to find the entailment between the comments. It also has a high computational cost since entailment has to be computed for all pairs of comments.
- The ranking strategies do not require manual intervention. The Top-k TextRank strategy was not adequate since it may return redundant summaries. However, TextRank combined with clustering and the centroid-based summarization algorithm tries to avoid redundant information and proved to be the recommended strategy.

To conclude, the major contributions of this chapter are several comment summarization approaches that essentially combine partitioning, clustering

and ranking with the centroid-based algorithm. To the best of our knowledge, no similar strategies have been proposed in the literature.

7
Conclusions

In this thesis, we created and evaluated sentiment analysis models of students' comments, and strategies to summarize students' comments. We tested the models and strategies using real data, obtained from (anonymized) student surveys applied at the Pontifical Catholic University of Rio de Janeiro from 2005 to 2021.

We presented two strategies for sentiment analysis in Chapter 5, summarized as follows:

1. The first strategy is based on a manually created dictionary that lists terms that represent the sentiment to be detected in the students' comments. We implemented this strategy as a tool, called *CourseObservatory*, which classifies the polarity of a set of students' comments and helps answer a set of questions that course and department coordinators may find useful.

2. The second strategy is based on the BERT language representation model. We tested three types of setup – zero-shot, from-scratch and fine-tuned – where the latter outperformed the other two. We showed that one may achieve good results without manually annotating a large number of comments. Because the model requires little training and is multilingual, it can be easily adapted to other universities.

We then applied the fine-tuned BERT model to predict the sentiment of the comments contained in the database since 2005.2. We analyzed in greater depth the results for 2019, 2020, and 2021 because we were interested in knowing how students reacted to the move to online classes forced by the COVID-19 pandemic. We may conclude that:

– Students acknowledged the effort of the teachers to keep classes running during 2020.1, and that the enthusiasm continued throughout 2020.2 and 2021.1.
– Students evaluated teacher 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.

We separated comment summarization into two problems and proposed different strategies, summarized as follows:

1. *Comment topic trending problem.* We proposed two strategies to detect the trending topics of a set of students' comments: Market-Basket Analysis and Topic Modeling. These strategies were inspired by tweet trending topics summarization techniques, but they explored the specific context of students' comments.

2. *Comment summarization problem.* We implemented three strategies to address this problem, used in isolation or in different combinations: partitioning, ranking, and entailment. The partitioning strategy was implemented using essentially syntactical approaches: grouping and clustering. The ranking strategy was implemented using TextRank. The direct application of TextRank proved not to be adequate, so it was combined with clustering and the centroid-based summarization algorithm. The entailment strategy was implemented using a specially trained BERT model. Again, the direct application of entailment proved not to be adequate, so it was combined with TextRank, clustering, and the centroid-based summarization algorithm. We recommend this last strategy because it returns a meaningful summary, does not need human intervention, and has a low computational cost.

In short, the major contributions of Chapter 6 were several comment summarization approaches that essentially combine partitioning, clustering and ranking with the centroid-based algorithm. To the best of our knowledge, no similar strategies have been proposed in the literature.

Although the strategies were motivated and tested with comments obtained from student survey data, all strategies can be applied to other scenarios. In the case of sentiment analysis, only new annotated data will be necessary to train the BERT-based model for other scenarios, but we saw that the set does not need to be large. For summarization, the TextRank strategy with clustering can be directly executed for any set of comments.

Lastly, the two sentiment analysis models were published in Jiménez et al. (2019) and Jiménez et al. (2021). We are now preparing an article about the comment summarization strategies to be submitted for publication.

As future work, we suggest to improve the sentiment analysis and comment summarization strategies to take into account additional student data, such as sex, age, zip code. We may also suggest expanding the experiments with the comment summarization strategies using other sets of manually created reference summaries, perhaps derived from the manual analysis of Section 6.3. One may also experiment with a committee of summarization algorithms to create a meaningful summary. Experiment with other datasets, outside of the

university scenario, are also desirable. Finally, we suggest testing other variants of the ROUGE metrics, or defining alternative metrics, especially those that would not require manually created reference summaries.

8
Bibliography

ABDIANSAH, A.; AZHARI, A.; SARI, A. Wertes: Web as external resources for textual entailment systems. **International Journal of Intelligent Engineering and Systems**, v. 11, p. 91–99, 2018. Cited in page 23.

AGRAWAL, R.; IMIELIńSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. **SIGMOD Rec.**, Association for Computing Machinery, New York, NY, USA, v. 22, n. 2, p. 207–216, jun. 1993. ISSN 0163-5808. Disponível em: <https://doi.org/10.1145/170036.170072>. Cited in page 62.

AHUJA, R. et al. The impact of features extraction on the sentiment analysis. **Procedia Computer Science**, v. 152, p. 341–348, 2019. Cited in page 19.

AKHTAR, M. S.; EKBAL, A.; BHATTACHARYYA, P. Aspect based sentiment analysis in Hindi: Resource creation and evaluation. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. European Language Resources Association (ELRA), 2016. p. 2703–2709. Disponível em: <https://www.aclweb.org/anthology/L16-1429>. Cited in page 20.

ALLAHYARI, M. et al. **Text Summarization Techniques: A Brief Survey**. 2017. Cited in page 24.

BALACHANDRAN, L.; KIRUPANANDA, A. 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)**, p. 1–7, 2017. Cited in page 21.

BALAHADIA, F. F.; FERNANDO, M. C. G.; JUANATAS, I. C. Teacher's performance evaluation tool using opinion mining with sentiment analysis. In: **2016 IEEE Region 10 Symposium (TENSYMP)**. [S.l.: s.n.], 2016. p. 95–98. Cited in page 20.

BANSAL, B.; SRIVASTAVA, S. Sentiment classification of online consumer reviews using word vector representations. **Procedia Computer Science**, v. 132, p. 1147–1153, 2018. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050918307610>. Cited in page 20.

BENTIVOGLI, L. et al. Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. **Language Resources and Evaluation**, v. 50, p. 95–124, 2016. Cited in page 25.

BHANDARI, M. et al. **Re-evaluating Evaluation in Text Summarization**. 2020. Cited in page 24.

BLAKE, C. The role of sentence structure in recognizing textual entailment. In: **Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing**. Prague: Association for Computational Linguistics, 2007. p. 101–106. Disponível em: <https://www.aclweb.org/anthology/W07-1417>. Cited in page 23.

CHATURVEDIA, I. et al. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. **Information Fusion**, v. 44, p. 65–77, December 2018. Cited in page 19.

DEVLIN, J. et al. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://www.aclweb.org/anthology/N19-1423>. Cited 4 times in pages 11, 22, 32, and 33.

FABBRI, A. R. et al. SummEval: Re-evaluating Summarization Evaluation. **Transactions of the Association for Computational Linguistics**, v. 9, p. 391–409, 04 2021. ISSN 2307-387X. Disponível em: <https://doi.org/10.1162/tacl_a_00373>. Cited in page 24.

GAO, Z. et al. Target-dependent sentiment classification with bert. **IEEE Access**, v. 7, p. 154290–154299, 2019. Cited in page 22.

GILYADOV, J. **Vector Representations of Words**. 2017. Disponível em: <https://israelg99.github.io/2017-03-22-Vector-Representations-of-Words/>. Cited in page 29.

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques, third edition**. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. ISBN 0123814790. Disponível em: <http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1>. Cited in page 62.

JIMéNEZ, H. G. et al. Sentiment analysis of student surveys - a case study assessing the impact of the covid-19 pandemic on higher education teaching. In: **14th International Conference on Educational Data Mining (EDM 2021), Paris, France, June 29 - July 2, 2021**. [s.n.], 2021. Disponível em: <https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_15.pdf>. Cited 5 times in pages 17, 21, 40, 54, and 86.

JIMéNEZ, H. G. et al. Courseobservatory: Sentiment analysis of comments in course surveys. In: **19th IEEE International Conference on Advanced Learning Technologies, ICALT 2019, Maceió, Brazil, July 15-18, 2019**. IEEE, 2019. p. 176–178. Disponível em: <https://doi.org/10.1109/ICALT.2019.00053>. Cited 5 times in pages 17, 21, 40, 53, and 86.

KHAN, S. **BERT, RoBERTa, DistilBERT, XLNet — which one to use?** 2019. Disponível em: <https://towardsdatascience.com/

bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8>. Cited in page 32.

KHOO, C. S.; JOHNKHAN, S. B. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. **Journal of Information Science**, v. 44, n. 4, p. 491–511, 2018. Cited in page 20.

KLEINDESSNER, M.; AWASTHI, P.; MORGENSTERN, J. Fair k-center clustering for data summarization. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). **Proceedings of the 36th International Conference on Machine Learning**. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 3448–3457. Disponível em: <https://proceedings.mlr.press/v97/kleindessner19a.html>. Cited in page 24.

LEMBERGER, P. Deep learning models for automatic summarization. **CoRR**, abs/2005.11988, 2020. Disponível em: <https://arxiv.org/abs/2005.11988>. Cited in page 24.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. **Mining of Massive Datasets**. 3. ed. USA: Cambridge University Press, 2020. ISBN 9781108476348. Cited 2 times in pages 29 and 31.

LI, J.; QIU, L. 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)**. [S.l.: s.n.], 2017. v. 1, p. 776–779. Cited in page 19.

LI, X. et al. Enhancing bert representation with context-aware embedding for aspect-based sentiment analysis. **IEEE Access**, v. 8, p. 46868–46876, 2020. Cited in page 22.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013>. Cited 2 times in pages 24 and 28.

LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012. Cited in page 19.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: _____. **Mining Text Data**. Boston, MA: Springer US, 2012. p. 415–463. ISBN 978-1-4614-3223-4. Disponível em: <https://doi.org/10.1007/978-1-4614-3223-4_13>. Cited in page 19.

LYTRAS, M. D. et al. Sentiment analysis to evaluate teaching performance. **Int. J. Knowl. Soc. Res.**, IGI Global, v. 7, n. 4, p. 86–107, October 2016. ISSN 1947-8429. Cited in page 21.

Lü, X. Statistical substring reduction in linear time. In: **Proceeding of the 1st International Joint Conference on Natural Language Processing (IJCNLP-2004). Volume 3248 of Lecture Notes in Computer Science**. [S.l.]: Springer, 2004. p. 320–327. Cited in page 42.

MAJUMDER, G. et al. Semantic Textual Similarity Methods, Tools, and Applications: A Survey. **Computación y Sistemas**, scielomx, v. 20, p. 647 – 665, 12 2016. ISSN 1405-5546. Disponível em: <http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462016000400647&nrm=iso>. Cited in page 23.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). **Proceedings of the 9th Python in Science Conference**. [S.l.: s.n.], 2010. p. 56 – 61. Cited in page 71.

MEANINGCLOUD. **Sentiment Analysis in Excel: getting started**. 2016. Disponível em: <https://www.meaningcloud.com/blog/sentiment-analysis-excel-getting-started>. Cited in page 42.

MENAHA, R. et al. Student feedback mining system using sentiment analysis. **International Journal of Computer Applications Technology and Research**, v. 6, n. 1, p. 51–55, 2017. Cited in page 21.

MIHALCEA, R.; TARAU, P. TextRank: Bringing order into text. In: **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 404–411. Disponível em: <https://aclanthology.org/W04-3252>. Cited in page 29.

NAIK, S. et al. Tweet summarization: A new approch. In: **2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)**. [S.l.: s.n.], 2018. p. 1022–1025. Cited in page 24.

NAZARE, S. P. et al. Sentiment analysis in twitter. **International Research Journal of Engineering and Technology (IRJET)**, v. 5, n. 1, p. 880–886, January 2018. Cited in page 19.

OLIVEIRA, D. N. de; MERSCHMANN, L. H. de C. Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in brazilian portuguese language. **Multimedia Tools and Applications**, Springer, p. 1–22, 2021. Cited in page 21.

PANG, B.; LEE, L. **Opinion mining and sentiment analysis (foundations and trends (R) in Information Retrieval)**. [S.l.]: Now Publishers Inc, 2008. Cited in page 19.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Cited in page 28.

POTA, M. et al. An effective bert-based pipeline for twitter sentiment analysis: A case study in italian. **Sensors**, v. 21, n. 1, 2021. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/21/1/133>. Cited in page 22.

PRUSA, J. D.; KHOSHGOFTAAR, T.; DITTMAN, D. Impact of feature selection techniques for tweet sentiment classification. In: **FLAIRS Conference**. [S.l.: s.n.], 2015. Cited in page 19.

REAL, L.; FONSECA, E.; OLIVEIRA, H. G. The assin 2 shared task: A quick overview. In: QUARESMA, P. et al. (Ed.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2020. p. 406–412. ISBN 978-3-030-41505-1. Cited 2 times in pages 23 and 75.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: **Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks**. Valletta, Malta: ELRA, 2010. p. 45–50. <http://is.muni.cz/publication/884893/en>. Cited 2 times in pages 29 and 71.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019. Disponível em: <https://arxiv.org/abs/1908.10084>. Cited 3 times in pages 23, 32, and 33.

ROSS, S. M. **Introduction to probability and statistics for engineers and scientists**. [S.l.]: Academic Press, 2020. Cited in page 49.

ROSSIELLO, G.; BASILE, P.; SEMERARO, G. Centroid-based text summarization through compositionality of word embeddings. In: **MultiLing@EACL**. [S.l.: s.n.], 2017. Cited 3 times in pages 24, 30, and 31.

SANTOS, C. L.; RITA, P.; GUERREIRO, J. Improving international attractiveness of higher education institutions based on text mining and sentiment analysis. **International Journal of Educational Management**, Emerald Publishing Limited, 2018. Cited in page 20.

SINDHU, I. et al. Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation. **IEEE Access**, IEEE, v. 7, p. 108729–108741, 2019. Cited in page 20.

SIVAKUMAR, M.; REDDY, U. S. Aspect based sentiment analysis of students opinion using machine learning techniques. In: IEEE. **2017 International Conference on Inventive Computing and Informatics (ICICI)**. [S.l.], 2017. p. 726–731. Cited in page 21.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. [S.l.: s.n.], 2020. Cited in page 66.

SOUZA, K. França de; PEREIRA, M. H. R.; DALIP, D. H. Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro. **Abakós**, v. 5, n. 2, p. 79–96, maio 2017. Disponível em: <http://periodicos.pucminas.br/index.php/abakos/article/view/P.2316-9451.2017v5n2p79>. Cited in page 21.

SOUZA, M.; VIEIRA, R. Sentiment analysis on twitter data for portuguese language. In: . [S.l.: s.n.], 2012. p. 241–247. ISBN 978-3-642-28884-5. Cited in page 21.

STEINBERGER, J.; JEZEK, K. Evaluation measures for text summarization. **Computing and Informatics**, v. 28, p. 251–275, 01 2009. Cited in page 24.

SUN, C.; HUANG, L.; QIU, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. **CoRR**, abs/1903.09588, 2019. Disponível em: <http://arxiv.org/abs/1903.09588>. Cited in page 22.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Addison Wesley, 2005. Hardcover. ISBN 0321321367. Disponível em: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0321321367>. Cited in page 62.

TATAR, D. et al. Textual entailment as a directional relation. **Journal of Research and Practice in Information Technology**, v. 41, 03 2009. Cited in page 23.

WAGNER, J. et al. The brwac corpus: A new open resource for brazilian portuguese. In: . [S.l.: s.n.], 2018. Cited in page 66.

ZHOU, J.; YE, J.-m. Sentiment analysis in education research: a review of journal publications. **Interactive Learning Environments**, Taylor & Francis, p. 1–13, 2020. Cited in page 20.

A
Questionnaire for In-Person Disciplines

Evaluate teacher performance:

1. Apresentação do programa no início do curso
2. Uso de recursos didáticos adequados ao conteúdo
3. Uso da bibliografia como apoio ao aprendizado
4. Cumprimento do programa do curso
5. Avaliação de aprendizagem dos alunos
   (a) Avaliação compatível com o conteúdo trabalhado
   (b) Avaliação com correção adequada
6. Atividades práticas
   (a) Orientação e supervisão da prática
   (b) Clareza na articulação da prática com a teoria
7. Clareza na exposição do conteúdo
8. Habilidade para motivar o interesse dos alunos
9. Incentivo à autonomia do aluno no processo de aprendizagem
10. Disponibilidade para tirar dúvidas
11. Assiduidade
12. Pontualidade
13. Relacionamento professor-aluno
14. Avaliação global do Professor
15. Comentários

Evaluate the discipline:

1. Clareza dos Tópicos do Programa
2. Bibliografia
   (a) Adequação do conteúdo ao programa

   (b) Quantidade de leitura compatível à carga horária

3. Equilíbrio entre o volume de tarefas (exercícios, programas computacionais, etc.) e a carga horária proposta

4. Atividades práticas

   (a) Articulação das atividades práticas com a teoria

   (b) Equilíbrio entre o volume de atividades práticas e teoria

5. Interesse do aluno pelos conteúdos oferecidos

6. Comentários

B
Questionnaire for Online Disciplines

Evaluate teacher performance:

1. Apresentação do programa no início do curso
2. Uso adequado dos recursos do ambiente virtual de aprendizagem
3. Uso adequado do material instrucional
4. Uso da bibliografia como apoio ao aprendizado
5. Cumprimento do programa do curso
6. Avaliação de aprendizagem dos alunos
   (a) Avaliação compatível com o conteúdo trabalhado
   (b) Avaliação com correção adequada
7. Atendimento ao aluno
   (a) Presteza
   (b) Qualidade
8. Habilidade para mediar a discussão dos conteúdos
9. Habilidade para motivar o interesse dos alunos
10. Incentivo à autonomia do aluno no processo de aprendizagem
11. Relacionamento professor-aluno
12. Avaliação global do Professor
13. Comentários

Evaluate the discipline:

1. Clareza dos Tópicos do Programa
2. Adequação do material instrucional ao conteúdo do programa
3. Bibliografia
   (a) Adequação do conteúdo ao programa
   (b) Quantidade de leitura compatível à carga horária

4. Equilíbrio entre o volume de tarefas (exercícios, programas computacionais, etc.) e a carga horária proposta

5. Interesse do aluno pelos conteúdos oferecidos

6. Comentários

C
Questionnaire for the Covid-19 Period

Evaluate teacher performance:

1. Adaptação do programa ao ambiente online
2. Uso adequado dos recursos e ambientes virtuais de aprendizagem
   (a) Ambientes Virtuais de Aprendizagem - AVA (Moodle, Maxwell, Google, etc.)
   (b) Plataformas para videoconferência (Zoom, etc.)
   (c) Autosau
   (d) Repositórios de arquivos na nuvem (drives)
   (e) Aplicativos de mensagens instantâneas (WhatsApp, Messenger, etc.)
3. Uso adequado do material instrucional (bibliografia, aulas gravadas, outros vídeos, links e materiais externos relevantes)
4. Cumprimento do programa originalmente proposto para o contexto presencial
5. Avaliação de aprendizagem dos alunos
   (a) Avaliação compatível com o conteúdo trabalhado
   (b) Avaliação com correção adequada
   (c) Avaliação com feedback adequado para o aluno
6. 6-Atendimento ao aluno
   (a) Presteza
   (b) Qualidade
7. Clareza na exposição do conteúdo
8. Capacidade de organização no contexto online
9. Habilidade para mediar a discussão dos conteúdos
10. Habilidade para motivar o interesse dos alunos
11. Incentivo à autonomia do aluno no processo de aprendizagem

12. Relacionamento professor-aluno
13. Avaliação global do Professor
14. Você gostaria de fazer algum comentário sobre o desempenho do professor?

Evaluate the discipline:

1. Clareza dos Tópicos do Programa
2. Adequação do material instrucional ao conteúdo do programa
3. Equilíbrio entre o volume de tarefas (exercícios, programas computacionais, etc.) e a carga horária proposta
4. Interesse do aluno pelos conteúdos oferecidos
5. Você gostaria de fazer algum comentário sobre pontos positivos ou facilidades da experiência de participar desta disciplina no ambiente online?
6. Você gostaria de fazer algum comentário sobre pontos negativos ou dificuldades da experiência de participar desta disciplina no ambiente online?
7. Você gostaria de fazer algum comentário adicional?

D
Entailment

Sentence A	Sentence B
admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom.	as aulas foram ótimas.
admiro o esforço e a humildade do professor para se adaptar ao atual recurso tecnológico de aulas através do aplicativo zoom.	atencioso, próximo ao aluno, didático.
as aulas foram ótimas.	atencioso, próximo ao aluno, didático.
excelente professor!	atencioso, próximo ao aluno, didático.
excelente professor!	excelente professor
excelente professor!	professor excelente!
excelente professor!	excelente.
excelente professor!	maravilhoso!
atencioso, próximo ao aluno, didático.	as aulas foram ótimas.
atencioso, próximo ao aluno, didático	excelente professor!
atencioso, próximo ao aluno, didático.	prestativo, e de uma competência e simplicidade extraordinária.
atencioso, próximo ao aluno, didático.	excelente professor
atencioso, próximo ao aluno, didático.	professor excelente!
atencioso, próximo ao aluno, didático	excelente.
atencioso, próximo ao aluno, didático.	maravilhoso!
prestativo, e de uma competência e simplicidade extraordinária.	as aulas foram ótimas.
prestativo, e de uma competência e simplicidade extraordinária.	excelente professor!
prestativo, e de uma competência e simplicidade extraordinária.	atencioso, próximo ao aluno, didático.
prestativo, e de uma competência e simplicidade extraordinária.	excelente professor
prestativo, e de uma competência e simplicidade extraordinária.	professor excelente!
prestativo, e de uma competência e simplicidade extraordinária.	excelente.
prestativo, e de uma competência e simplicidade extraordinária.	maravilhoso!
excelente professor	excelente professor!
excelente professor	atencioso, próximo ao aluno, didático.
excelente professor	professor excelente!
excelente professor	excelente
professor excelente!	excelente professor!
professor excelente!	atencioso, próximo ao aluno, didático.
professor excelente!	excelente professor
professor excelente!	excelente.
excelente.	excelente professor!
excelente.	atencioso, próximo ao aluno, didático.
excelente.	excelente professor
excelente.	professor excelente!
maravilhoso!	excelente professor!
maravilhoso!	atencioso, próximo ao aluno, didático.
maravilhoso!	professor excelente!
maravilhoso!	excelente.