BSc Thesis Psychology

Text Mining to Improve Education: An Evaluation of Text Mining in the Student Feedback Process

Christian T. Claessen

Dr G. Sedrakyan (1st Supervisor)
Dr M. Amir Haeri (2nd Supervisor)

University of Twente, The Netherlands
Department of Behavioural and Social Sciences

June 30, 2022

UNIVERSITY OF TWENTE.

Abstract

Background: When it comes to evaluating education, student feedback surveys are common practice. Students are then typically asked to answer both open and Likert-type questions. While there is a consensus on the richness of qualitative data, its analysis is costly. One solution that would allow universities to take full advantage of their data, is text mining.
Objective: This study investigated the usefulness of text mining methods for the analysis of student feedback by applying two of those techniques, namely sentiment analysis and text summarization, as well as by mapping out the different state-of-the-art techniques applied to student feedback.
Methods: Student survey reports from 5 courses at the University of Twente between 2019 and 2022 were gathered. Each survey consisted of 9 sections with a Likert-type average and comments for each of the segments. First, a BERT (Bidirectional Encoder Representations from Transformers) sentiment analysis that produced a score from 1 to 5 was conducted on all of the comments and a mean of those sentiment scores was calculated for each section. These means of the sentiment analysis were then correlated with the Likert-Type data. Secondly, 140 comments on one of the five surveys were manually coded for polarity and then correlated with their respective sentiment scores. A PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) text summarizer was used for the same survey data. The produced summaries were then evaluated by a human assessor. At last, a systematic literature review was conducted and led to the analysis of ten research articles.
Results: Firstly, it was found that most articles focused on sentiment analysis and/or clustering, being followed closely by categorization. No other methods were identified in the literature study. Secondly, a moderate to strong relationship was found for the sentiment analysis ($r = .612$, $p < 0.001$). Third, sentiment scores and manually coded polarity judgements showed non-normality, therefore a Spearman correlation coefficient was calculated. This showed that the two variables were strongly correlated ($r = .805$, $p < 0.001$). For the evaluation of the text summarization, a mean human assessment of $M=1.82$ with a standard deviation $SD=.21$ was found.
Conclusion: Sentiment analysis may be a useful tool for replacing numerical measurements of student feedback. Text summarization on individual comments did not yield promising results, which may be due to the shortness of comments. Moreover, current literature mainly focuses on sentiment analysis, clustering, and categorization. Consequently, future efforts may expand this research by using larger samples and applying different methods.

Contents

Introduction

Gathering student feedback has long been established as a common practice of quality assurance in higher education institutions. One way that universities approach this is by making use of standardized evaluation questionnaires which are carried out after students have completed a specific component of their education. Questionnaires then usually contain both Likert-type and open-ended questions/commentary sections. The goal of these questionnaires can be summarized into three points: Firstly, to gather feedback on the effectiveness of teaching for both teachers and administrators, secondly to inform students when making course-selection decisions and lastly to get data for educational research (Marsh & Dunkin, 1992, as cited in Richardson, 2005).

Therefore, most feedback questionnaires contain two distinct types of data: Firstly, quantitative data is gathered in the form of Likert scale questions. Secondly, qualitative data gets queried in the form of general comments or text answers to open-ended questions. In general, research indicates that students' comments offer more detailed and specified information about students' opinions in comparison to quantitative measures, but often lack systematic analysis (Scott, 2021). While quantitative data allows decision-makers to observe broad-scale trends, comments provide a richer and more detailed image of students' perceptions (Mandouit, 2018; Scott, 2021). According to Scott (2021), quantitative feedback lacks information on the importance and weight given to specific measurements. Moreover, asking for written feedback ensures that items on questionnaires are complete and represent students' interests (Scott, 2021). Generally, textual feedback gives students an opportunity to voice their opinions, ideas and suggestions, thus representing a valuable resource for the improvement and development of universities (Brockx et al., 2012; Palmer & Campbell, 2013).

The consequence of this demand for textual feedback paired with the organizational constraints of higher educational institutions creates a gap for new ideas in processing feedback data. Novel technologies in the realm of data analysis present one way of approaching the practical limitations of educational institutions. Text mining in particular has been one domain of methodologies, that has become vastly popular in the past years, also within the educational sector (Ferreira-Mello et al., 2019; Shah & Pabel, 2019). Hence, a significant amount of research has explored the possibilities of automating textual feedback analysis in recent years (Scott, 2021; Shah & Pabel, 2019; Ulfa et al., 2020).

Objective and Research Question

Despite the previous attempts in automating parts of the analysis of qualitative student feedback, prior research has mostly focused on data science methodologies, without regarding how these methods may be useful in improving education. Consequently, most of previous research has failed to show how these text mining methodologies can help in the analysis of student feedback. The goal of this research is therefore to identify how text mining is useful in analysing student feedback. Thus, this paper aims to answer the following research questions: ***How can text mining assist educational institutions in the process of analysis of student evaluations of teaching?***

To further define the scope of this research, the following three questions will be answered:

1. How can sentiment analysis support the process of analysing student evaluations of teaching?
2. How can text summarization support the process of analysing student evaluations of teaching?
3. What state-of-the-art text analytics techniques are currently employed in the analysis of student feedback and how can they support this analysis?

Related Work

In this section, the current state of the literature on the use of text mining in analysing student feedback will be discussed. In today’s modern marketplace of higher education, universities must take students’ feedback serious for several reasons. First and most importantly, student feedback should be used to improve education. Therefore, feedback serves the function of showcasing what areas of education require improvement and how this improvement may come about (Harvey, 2003; Marsh & Dunkin, 1992, as cited in Richardson, 2005). Related to this first objective, the second goal of feedback is to guide, monitor and evaluate any attempt to improve teaching and hence the student's learning process. (Harvey, 2003). Harvey (2003) refers to this as benchmarking of teaching quality. Therefore, any increase or decrease in students’ evaluations can provide insight into the improvement process. Importantly, student feedback provides an insider's perspective and thereby relays information unavailable to administrators or teachers (Mandouit, 2018).

Textual Student Feedback

The value of textual student feedback in improving education has most definitely been acknowledged in the literature. However, research also points out some significant resource limitations associated with the analysis of textual student feedback (Santhanam et al. 2018; Shah & Pabel, 2019). These limitations typically refer to universities' limited resources in conducting meaningful feedback analyses, as well as the overall low reliability of manual assessment (Richardson, 2005; Rowley, 2003; Shah & Pabel, 2019). Moreover, these constraints are amplified by the increasing number of students that universities have to facilitate, which are then in turn leading to a larger number of responses. Additionally, the fact that universities often simply return textual feedback to teachers without providing any structure, further limits its effectiveness (Kember et al., 2002). Overall, these limitations create a dilemma: Universities would like to make use of the qualitative data they gather, but ultimately lack the resources. Resulting from this is the need for new innovations in the analysis of qualitative student feedback.

Innovations in Text Data Processing

Making use of large amounts of data while being resource-aware is not a new problem within the world of education. The recent growth of online education has caused a massive increase in qualitative data availability, which in turn has incentivized universities to invest in a variety of new data processing tools, including text mining (Ferreira-Mello et al., 2019; Romero & Ventura, 2020; Scott, 2021). Many educational tasks require the production of largely unstructured text data, such as essays, open questions, or forum discussions. Consequently, there have also been previous attempts in analysing qualitative data from student evaluations of teaching (Ferreira-Mello et al., 2019). This section of the paper will provide an overview of four different text mining techniques that have been previously used on student feedback. The methods of clustering and categorization will be explored first, after which the text summarization and sentiment analysis will be regarded more in-depth. Moreover, at least one example application is given for each of these methods.

Categorization and Clustering

Two of the most widely used text mining techniques within the educational domain are clustering and categorization (Ferreira-Mello et al., 2019; Gaikwad et al., 2014). While clustering is a method that can be used to group different documents with similar content together, categorization is used to classify text documents with one or more pre-defined features. Categorization consequently differs from clustering in so far as categories are pre-defined (Gaikwad et al., 2014).

Shah & Pabel (2019) provide a good example of how clustering may be useful in giving evaluators a first overview of large qualitative datasets. In their study, they made use of the text analytics software Leximancer to compare written feedback of online and offline students. Leximancer thereby identifies frequently occurring concepts and automatically groups related concepts into larger categories. The outcome of that analysis is then visualized in a concept map. Shah & Pabel (2019) ultimately conclude that Leximancer is a useful tool for universities and other stakeholders in giving structure to large volumes of commentary data.

One of the first successful and large-scale attempts in implementing categorization has been Australia's *CEQuery* project (Scott, 2021). CEQuery was a qualitative data analysis tool for the Australian course experience questionnaire, which is Australia's national benchmarking survey completed by university students after graduation (Scott, 2021). CEQuery used a dictionary with educational contexts to search through and subsequently classify comments into five categories. Moreover, CEQuery has widely been viewed as a success, since it allowed universities and researchers to identify important improvement points in education, while also proving the concept of text mining in qualitative educational feedback.

Text Summarization

Another text mining solution for reducing the constraints of qualitative data is text summarization. Text summarization automatically shortens documents down to the most relevant and essential information (Ferreira-Mello et al., 2019; Gaikwad et al., 2014). According to Gaikwad et al. (2014), this allows users to evaluate whether a lengthy text document contains valuable information and should hence be read in its entirety, thereby saving valuable time. Although being vastly less popular than the previous methods, text summarization has been adopted for a wide variety of educational purposes such as providing students with writing assistance or evaluating online learning platform posts (Ferreira-Mello et al., 2018). Regarding the evaluation of student feedback research is scarce (Ferreira-Mello et al., 2018). One of the few approaches comes from Luo et al. (2016) who built a summarization system that successfully identifies and extracts important phrases from a body of students' comments. Overall however, text summarization seems to hold great potential for more inquiries.

Sentiment Analysis

Lastly, sentiment analysis, also referred to as opinion mining, is a data mining tool that enables users to identify states of affection as expressed in the textual content. Sentiment analysis is used to understand and rate the polarity, emotionality, or sentiment in textual contents (Ulfa et al., 2020). Ratings are then usually expressed as positive or negative or can alternatively also be expressed in scales (Nasim et al., 2017).

According to Ferreira-Mello et al. (2019), there is already a good basis of research on sentiment analysis in the educational domain, however, its full potential has not been explored yet. This conclusion does not only hold for the field of education at large but is also true for sentiment analysis on student feedback. To be more specific, there have been a number of papers applying sentiment analysis in this context (Aung & Myo, 2017; Rani & Kumar, 2017; Sadriu et al., 2022; Toçoğlu and Onan, 2020; Ulfa et al., 2020,). Nasim and colleagues (2017) for example successfully combined machine learning with lexicon-based approaches of sentiment analysis on qualitative student feedback. In the end, this algorithm managed to mimic a manual coder in 93% of all cases. In a different study, Onan (2019) showcased the sophistication of deep learning sentiment analysis in student feedback comments with an accuracy of 98.29%. Despite the success of these studies, most research on student feedback sentiment analysis has come from a computer science perspective. Hence, most research in the past has shown the accuracy of sentiment analysis, without considering the implications or usage of sentiment information. These are first, the connection between sentiment scores and quantitative measurements, as well as secondly students' emotional states and opinions (Dunlosky et al., 2013).

Firstly, sentiment scores might be correlated with corresponding numerical measures (Neumann & Linzmayer, 2021). This indicates the redundancy of numerical measures by the novel method of sentiment analysis. In other words, using sentiment analysis on written might render numerical measures irrelevant. Although this has been shown within experimental environments such as in Neumann & Linzmayer (2021), it is not clear whether these results are also generalizable to formal student feedback with Likert scales.

Secondly, sentiment Analysis has proven to be a reliable tool for understanding a writer's emotions (Neumann & Linzmayer, 2021). Understanding emotions in the context of educational feedback is advantageous because comprehending students' emotions gives universities and teachers the ability to adapt teaching environments for the facilitation of more positive ones. This is important, because of the key role that emotions play within education. Emotions are strongly associated with the cognitive processes involved in learning, by regulating attention, encoding, retrieval, and problem-solving (Tyng et al., 2017). Moreover, research has shown positive emotions to be strongly associated with academic success, as well as students' motivation to learn (Loderer et al., 2020; Mega et al., 2014). Importantly, Pekrun (2006) adds that emotional states are strongly tied to a student's learning environment, therefore suggesting that educational institutions should design learning environments (including courses, teaching methods, and examinations) in such a manner as to prevent malicious emotions and promote healthy ones (Pekrun, 2006). For educators, this means that obtaining information about these emotions can be fruitful. Sentiment analysis consequently is a technique that can assist in obtaining insights into students' affective states and acting upon them.,

Summing up, universities use qualitative student feedback in order to gain an insider's perspective on the issues arising within their classrooms. While textual feedback is an undeniably valuable resource, universities are often not able to take full advantage of this data. The reason for that is that the traditional analysis of student feedback by hand is costly and inefficient, a problem that is made even worse by the fact that qualitative data is commonly returned to teachers without structure. One possible solution for this might come from novel text mining techniques, which automatically extract information from text documents and therefore reduce the need to manually process the entirety of it. These techniques include categorization, clustering, text summarization and sentiment analysis all of which have been applied in the educational domain already. Nevertheless, especially research on the last two methods, summarization, and sentiment analysis, has failed to integrate technology with an account for how the produced data might be useful for feedback evaluators. Furthermore, there have been only a handful of studies examining the utilization of text summarization on student feedback. This research will therefore attempt to fill these gaps of research.

Aims of the Research

The goal of this study was to evaluate the usage of text mining techniques in understanding qualitative student feedback. To narrow the scope of this research, two text mining techniques were opted for, namely sentiment analysis and text summarization. These methods were chosen, as research on those specific techniques so far has not fully mapped out their potential yet with regard to textual feedback in education, as will be explained further in this chapter. Additionally, the usage of other possible methodologies was explored in a systematic literature review. Consequently, this study consisted of three major parts.

Firstly, the objective in applying sentiment analysis was to understand whether sentiment scores could serve as a replacement of Likert-type data. Such an approach would consequently allow students to simplify student feedback surveys to only open-ended questions. Another aim of applying sentiment analysis was to assess its capabilities in understanding emotions in comparison to a human. Secondly, one objective of this research was to investigate the use of text summarization on commentary data. No previous research has, to the best of our knowledge, attempted to use text summarization on individual commentary data. The goal of this paper was therefore to clarify whether such a technique could be effective for practitioners in reducing the amount of qualitative data to be analysed, by reducing the word count in students’ comments. Finally, this study aimed at providing a systematic overview of the methods currently used for student feedback analysis with the goal to reflect on complementary capabilities in the field of feedback analytics. This will enable researchers to clearly guide their future efforts towards the less well-adopted methodologies of text mining, and thereby fully take advantage of these new data mining tools.

Methods

The overall objective of this research was to identify how new data processing techniques can assist evaluators of qualitative student feedback in drawing meaningful conclusions from feedback reports. As stated, previous research on both sentiment analysis and text summarization has mostly focused on methodologies. The added value of this research therefore was to connect the possibilities of these methods with the potential informational gain for practitioners. Moreover, no previous study has compared sentiment analysis data with the results of Likert-scale data. Firstly, a sentiment analysis was conducted using the Bidirectional Encoder Representation for Transformers (BERT) method, while secondly a text summarization was done on a Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) model. The rationale of the choice of these specific methods is further explained in the sections below. Finally, to answer the last research question, a literature review on previous research on text mining on student feedback was conducted.

Sentiment Analysis

The Bidirectional Encoder Representation for Transformers (BERT) method with a pretrained model was chosen to perform the sentiment analysis on student feedback because BERT models are trained unsupervised on large quantities of data with the advantage of great generalizability onto many tasks (Devlin et al., 2018). This decision was made, due to the previous success of this method in different studies, as well as due to the limited amount of qualitative data available for training our own model (Cen et al., 2021, Mathew & Bindu, 2020). By using BERT this study follows the suggestions of Kastrati et al. (2021), who proposed its application for the task of student feedback analysis. Consequently, a pretrained BERT model was selected that was previously trained on more than 150 thousand product reviews (NLP Town, n.d.). Furthermore, this particular BERT model was selected, because of its ability to rate polarity on a scale of one to five, consequently being easily comparable with the questionnaire Likert-scale data.

Text Summarization

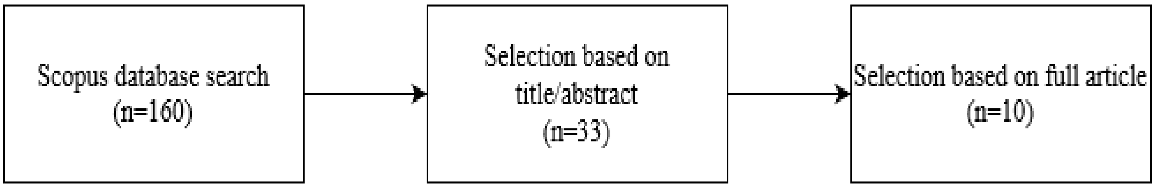
For the performance of the text summarization there were two possible methodological options, as text summarizers can be distinguished into extractive and abstractive systems (Ferreira-Mello, 2019). Extractive systems create their summary by identifying the most important sentences and extracting them word-by-word (Ferreira-Mello, 2019). In turn, an abstractive summarizer writes a brief version of a text document by generating novel sentences, or rephrasing old (Ferreira-Mello, 2019). For this study, we opted for an abstractive summarization system, since student feedback is often short in itself already, meaning that an extractive solution would likely omit meaningful sentences for its summary. In order to perform the abstractive summarization, a PEGASUS model was chosen. Developed by Google researchers in 2020, Pegasus is a pre-trained text summarization text mining tool, that transforms text data into short summaries (Zhang et al., 2020). The decision to use PEGASUS was made, because PEGASUS does not require any data for training (Google, n.d.).

Systematic Literature Review

Next, a systematic review of previous literature on text mining in education was conducted. The electronic literature database Scopus was selected as the main source of literature. Therefore, a search string was created. Since the goal of this study was to explore text mining methods, the first set of terms was “natural language processing”, “NLP”, “data processing”, “text mining”, “categorization”, “clustering”, “sentiment analysis”, and “summarization”. These terms were chosen since they cover the spectrum of available text mining techniques widely. Secondly, the term “student feedback” was selected, to specify the search for student feedback. Out of these terms, the following search string emerged: “ TITLE-ABS-KEY ((*"NLP"* OR *"natural language processing"* OR *"text mining"* OR *categorization* OR *clustering* OR *"sentiment analysis"* OR *summarization*) AND *"student feedback"*). This search string resulted in a total of 160 articles (Figure 1).

Figure 1.

Search flow



Study Selection

To select relevant literature, titles and abstracts of all found articles were screened and then either kept or omitted in accordance with the following exclusion criteria:

- study is not focused on higher education
- study does focus on Massive Open Online Course (MOOC) feedback
- duplicates
- research is solely methodologically/technologically oriented

Firstly, the goal of this study was to understand what methods had been applied in universities, therefore all studies that did not focus on higher education were omitted. Similarly, all articles analysing feedback on massive open online courses were excluded, since the nature of these courses is vastly different from actual university courses. Next, in order to be able to see what methods have been evaluated on their usefulness for practitioners, studies that solely focused on text mining methodologies were excluded. Lastly, all duplicates were removed. After exclusion, 33 articles were left, which were then narrowed down by applying three inclusion criteria:

- study uses formally gathered student feedback
- study uses an evaluation form
- goal of the study is to analyse the use of the method

For this step, the full texts of all articles were read and examined. Since this research is focused on text mining in student feedback questionnaires, only studies relating to student feedback gathered in a formal feedback evaluation questionnaire were included. Additionally, the remaining articles had to be designed with text mining as a major component of the study. In total, ten articles were included in the literature review, which are outlined in Appendix C.

Implementation

Within this section of the paper, I will address the concrete implementation of the two text mining systems, as well as their evaluation metrics. In general, there were two objectives for the implementation. Firstly, to predict the quantitative data in a student feedback survey with a sentiment analysis on the corresponding qualitative data. Secondly, to provide accurate summaries of the comment data. To address these goals, I used the data from a formal student feedback survey, which contained both qualitative and quantitative questions. The implementation of both systems was done in the programming language Python (see Appendix A & Appendix B) and can generally be split into three parts: (1) Pre-Processing the Data; (2) performing sentiment analysis/text summarization and measuring evaluation metrics (3) analysing the data statistically. This entire process is illustrated in Figure 2 for sentiment analysis and Figure 3 for text summarization.

Figure 2.

Three-Step Process for Sentiment Analysis

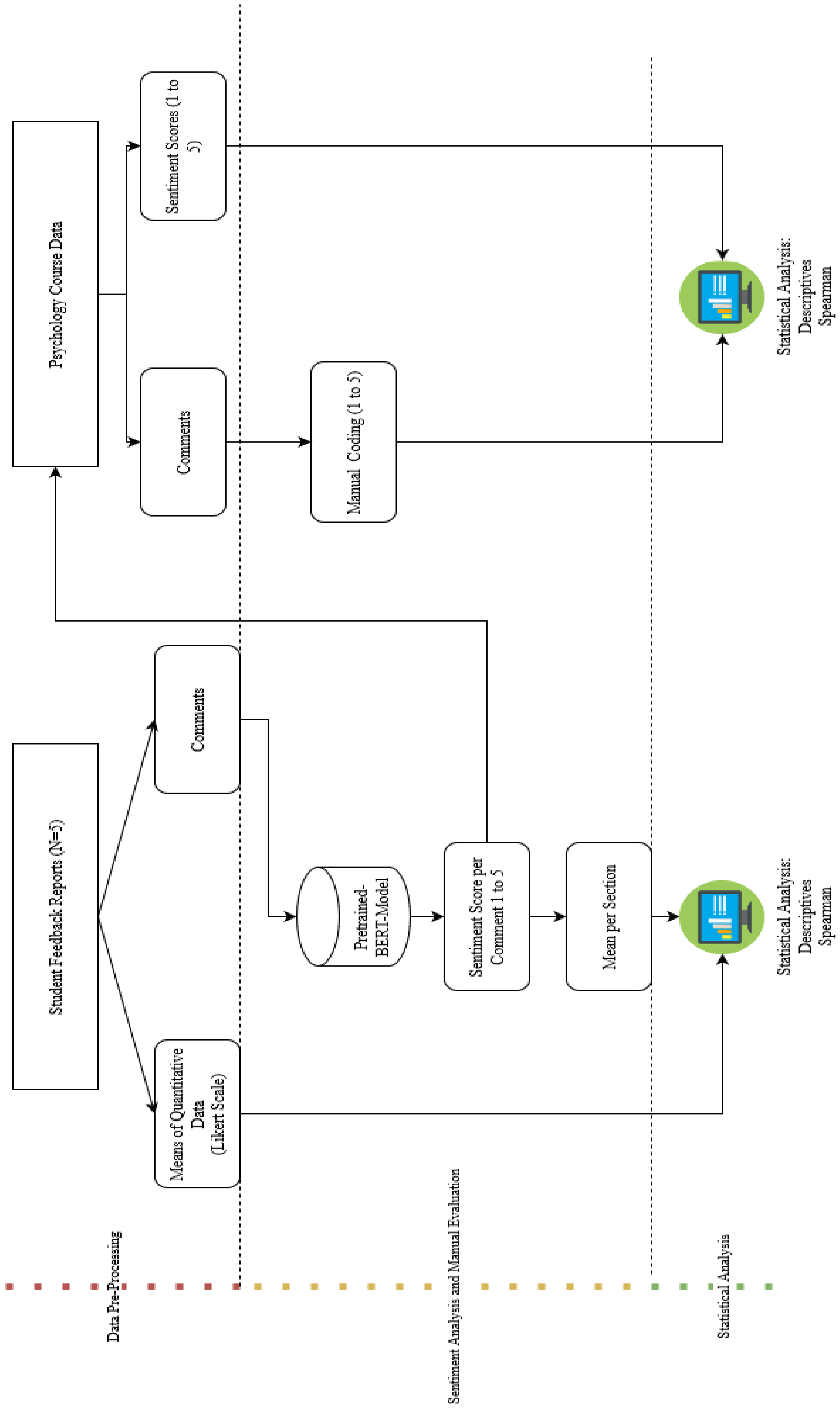
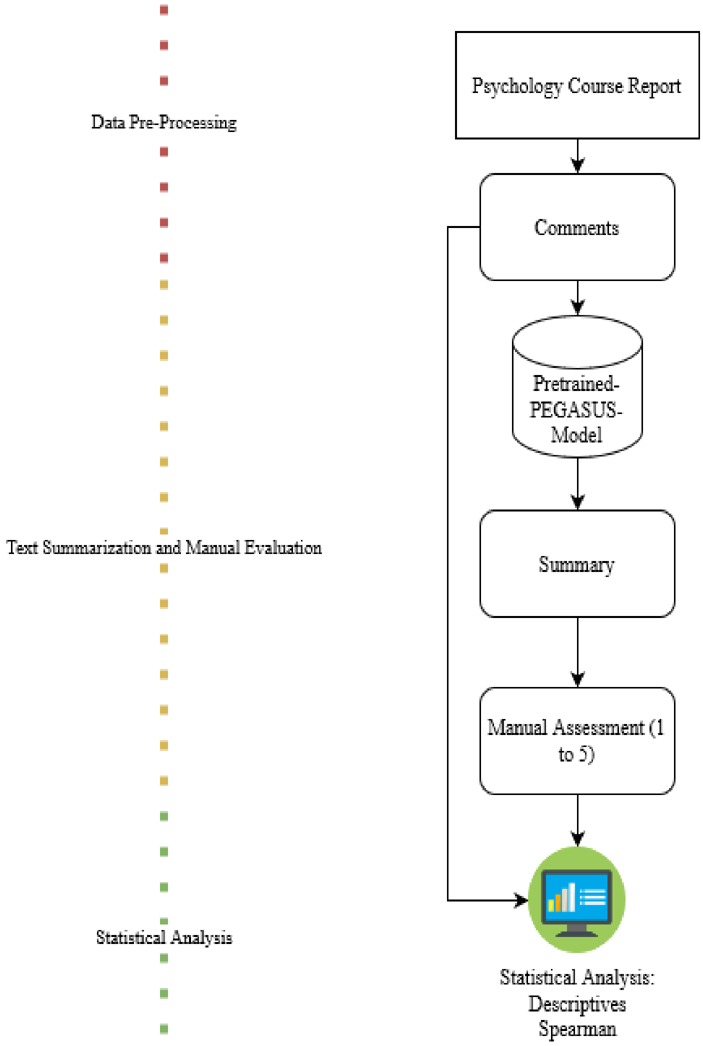

Data Pre-Processing
Sentiment Analysis and Manual Evaluation
Statistical Analysis
Student Feedback Reports (N=5)
Comments
Means of Quantitative Data (Likert Scale)
Pretrained-BERT-Model
Sentiment Score per Comment 1 to 5
Mean per Section
Statistical Analysis: Descriptives Spearman
Psychology Course Data
Comments
Sentiment Scores (1 to 5)
Manual Coding (1 to 5)
Statistical Analysis: Descriptives Spearman

Figure 3.

Three-Step Process for Text Summarization


Data Pre-Processing
Text Summarization and Manual Evaluation
Statistical Analysis
Psychology Course Report
Comments
Pretrained-PEGASUS-Model
Summary
Manual Assessment (1 to 5)
Statistical Analysis: Descriptives Spearman

Data

All in all, the entire data included five different course evaluation surveys conducted at the University of Twente. Students have filled out these questionnaires digitally in the years ranging from 2018 to 2022. The survey was structured into nine sections, in all except one of which students were asked different five-point Likert-scaled questions (Appendix D). Moreover, each section also contained space for comments, meaning that comments were already pre-grouped according to specific topics. These sections were: Module (1), Learning-1 (2), Learning-2 (3), Teaching (4) Project (5), Assessment (6), Study Load (7), Online Education (8), and strengths/points of improvement/appreciation (9).

Data Pre-Processing

Before conducting the two text mining methods, the data was sighted for any irregularities. Since some questionnaires differed in that they did not contain sections eight and nine, these were excluded from the data. Additionally, some sections contained more than one qualitative question but listed only one mean for all Likert-Type questions. The text answers to all these qualitative questions were therefore grouped. Section 7 for example contained both the question “Please explain why the study load of this module was not well-spread over the quartile” as well as “Please explain your neutral or positive answer(s) about the question block Study load”. All answers to these questions were regarded as answers to section 7. Moreover, all comments were extracted from each pdf-report and manually copied into three corresponding Microsoft Excel spreadsheets. The comments were then categorized into the specific sections they belonged to.

Sentiment Analysis

To perform the sentiment analysis, each comment first had to be tokenized using WordPiece (Devlin et al., 2018). To get a preliminary idea of sentiment analysis’ efficacy, three exemplary sentences were first coded manually on a scale from one to five and then compared to the BERT-model score (Table 1). After that first sentiment analysis yielded a score similar to that of manual coding, each token sequence was scored for polarity using the pre-trained BERT model (Devlin et al., 2018). Therefore, a BERT-model sentiment score was computed for each comment. Since the original data only contained means of the quantitative data, it was decided to compare only the mean sentiment of comments per section with the mean quantitative score. As a result, all means of sentiment scores were calculated per segment, thus allowing the comparison of the averages of the sentiment analysis with the averages of the quantitative survey part.

To understand how well a sentiment analysis can measure emotions in student feedback compared to a human, one of the five-course evaluation surveys was coded manually. Therefore, all 140 comments from a psychology course evaluation were first read and then assessed for polarity on a scale of one to five. These manually coded scores were then inserted along with the already calculated sentiment score of each comment into a Microsoft Excel sheet. The limits of this method are that a manual assessment is subjective and can therefore hold biases. The entire process is again illustrated in Figure 1.

Table 1.

Example of Sentiment Analysis using Likert-scale

Sentence	Sentiment Score	Manually Coded Score
"*It was well organized and good coherence*"	4	4
"*At times i feel like there wasn't enough guidance for the projects and the help with mendix was insufficient*"	2	2
"*During this module I learned a lot by myself for the project and I don't feel the teachers were ready for this kind of complex projects.*"	3	2

Text Summarization

After conducting the sentiment analysis, the second text mining technique, namely text summarization, was explored using the PEGASUS method. This was once again done only for the psychology course feedback. For PEGASUS summarization, the comments were first tokenized using the SentencePiece method as suggested by Zhang et al. (2020). Afterwards, each comment was summarized, and the summary was then saved alongside the original comment in a Microsoft Excel spreadsheet. Since two comments produced an error in the summarization system, these were excluded from the data. Next, to compare the differences in length, word counts were calculated for comments and summaries. As an evaluation metric for the summaries, the subjective assessment method as proposed in Beke and Szaszák (2016) was adopted. Therefore, all summaries were rated based on the question: "How well does the system summarize the narrated content in your opinion?" (Beke & Szaszák, 2016). This evaluation was once again performed manually by one assessor with the use of a five-point Likert-scale.

Statistical Analysis

Lastly, to evaluate the two text mining methods, three statistical analyses were performed within the statistical software SPSS 25.0. First, the comparison of sentiment scores with Likert-Type data was made, as well as a comparison within the psychology course evaluation data (manual coding v. sentiment analysis). For this analysis, descriptive statistics, means and standard deviations were calculated for sentiment and quantitative scores in order to get a preliminary idea of the relationship between the two variables. Next, a scatterplot of these scores was conducted for the same reason. Since both variables contained mean scores, a non-parametric test, the Spearman correlation coefficient, was conducted. For the interpretation of results, Akoglu (2018) suggests that significant correlations r >.70 can be considered as strong correlations. Moreover, a relationship can be named moderate for a p value between .40 and .60 (Akoglu, 2018).

For the second analysis of the psychology course data, the manual coding scores were compared with the sentiment analysis scores. Therefore, descriptive statistics, means and standard deviations for sentiment scores and manual scores were again calculated. Since, this time the data was purely discrete (1 to 5), a bar chart of means of sentiment and manual coding scores for each survey section was drawn. Furthermore, assumptions of normality were checked and on its basis, a Spearman correlation was calculated. Once again, Akoglu's (2018) criteria for labelling a correlation were applied again.

The third and last analysis of the text summarization performance was conducted based on a manual Likert-scale evaluation. Therefore, descriptive statistics, such as means and standard deviations were used. As an evaluation criterion, I adopted the wording of the Likert scale, ranging from 1 to 5: "Poor, Moderate, Acceptable, Good, Excellent" (Beke & Szaszák, 2016).

Systematic Literature Review

The aim of this literature review was to discover what different text mining methods have been researched in the context of analysing student feedback. As stated in the recent work section, four major text mining methods were identified, namely clustering, categorization, sentiment analysis, and text summarization. Therefore, ten articles were included in this review, all of which aimed at the discussion of different techniques and their usage on student feedback (Appendix C)

Based on the ten articles that had been reviewed, the two most used methods for analysing student feedback were identified to be sentiment analysis and clustering, each of them being included in five of the ten studies. In general, nine out of the ten studies included either a sentiment analysis, a clustering methodology, or both in their research (Bhaduri et al., 2021; Gottipati et al., 2017; Gronberg et al., 2021; Hynninen et al., 2020; Katz et al., 2021; Neumann & Linzmayer, 2021; Nitin et al., 2015; Shah & Pabel, 2019). These two methodologies were closely followed by categorization which had been included in four of the ten studies (Gottipati et al., 2017; Gottipatti et al., 2018; Nawaz et al., 2022; Nitin et al., 2015). Interestingly, none of the articles applied text summarization to comments. Similarly, no additional techniques were found.

Results

Furthermore, the results of the statistical analysis of the two applied methods, sentiment analysis and text summarization will be reported.

Sentiment Analysis

After excluding all incomplete and all non-English comments due to the language restriction of the BERT model, the total number of comments was 480. The number of comments per section ranged from 1 to 55 comments with a mean amount of M=13.71 (SD=10.42). Next, the minimum obtained mean of a quantitative (Likert-type) questionnaire section was 1.70, while the maximum score was 4.10 with a mean quantitative rating for all sections of M=3.42 (SD= .45). For the sentiment analysis, the scores ranged from 2.30 to 4.10 with the mean sentiment score of all sections being M=2.97 (SD= .55). Similarly, looking at the mean scores per questionnaire section, the quantitative mean score on average was higher than the sentiment score (Appendix E). On average the “Learning-2” section received the best ratings for both the sentiment analysis (M=3.60; SD=.45) and the quantitative scores (M=3.70; SD=.27).

Table 2.

Descriptive Statistics

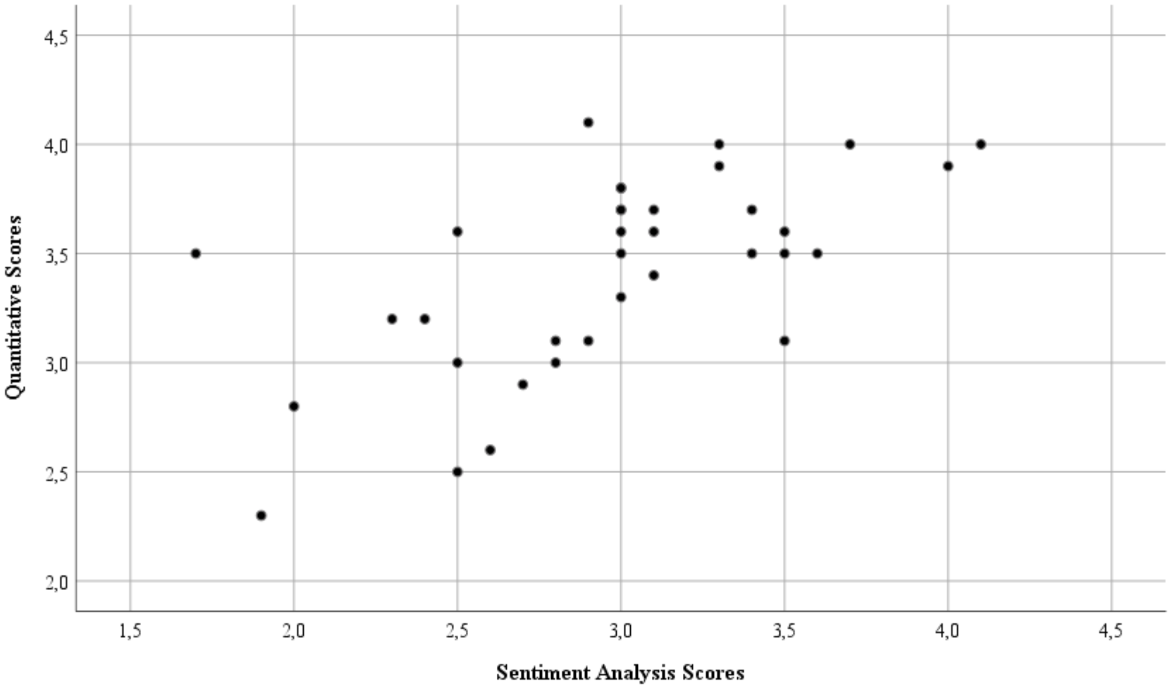
Variable	*M*	*SD*	Minimum	Maximum
1: Mean Quantitative Score	3.42	.45	1.70	4.10
2: Mean Sentiment Score	2.97	.55	2.30	4.10

Relationship between Quantitative Data and Sentiment Scores

In order to provide an answer to the first research aim on using sentiment analysis to predict topic-based Likert-scale scores, it was tested whether the means of the sentiment analysis correlate with the means of the quantitative measures. Plotting the scores obtained from the sentiment analysis against the quantitative rating reveals a positive relationship between sentiment and student ratings (Figure 4). Moreover, a Spearman Correlation Coefficient was calculated and demonstrated a significant correlation r = .612, p < .001. From this, it can be inferred that there is a significant correlation between quantitative data and sentiment scores. In addition, in line with Akoglu's (2018) User's guide to correlation criteria, these results indicate a moderate to a strong relationship between sentiment scores and quantitative scores.

Figure 4.

Scatterplot of the quantitative scores and the sentiment analysis scores

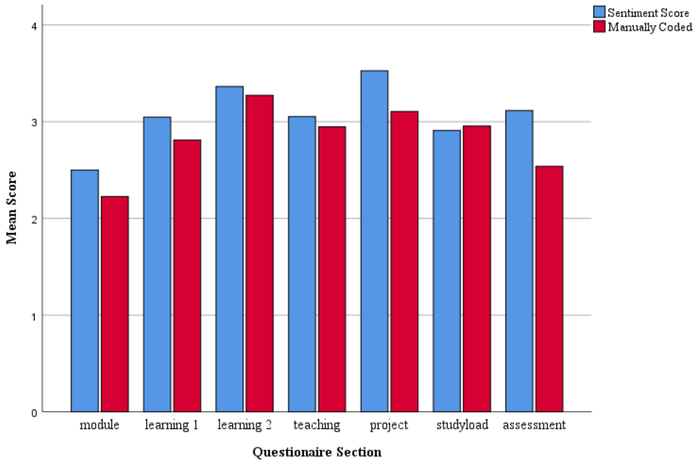


Relationship between Manual Coding and Sentiment Scores

Next, the dataset of the psychology course was analysed for differences in the manual coding and the sentiment score. The total number of comments was 140. In order to check the normality of data, a Shapiro Wilk test was used to check the assumption of normally distributed residuals. The value of p < .000 for manually coded scores, as well as the value of p < .000 for the sentiment scores, violated the assumption of normality. As a consequence, a Spearman correlation coefficient was calculated. The results indicated that scores obtained from the sentiment analysis were indeed highly correlated with manual scoring r = .805, p < 0.001. Following Akoglu (2018) this correlation revealed a strong relationship between manual coding and sentiment analysis scores.

Figure 5.

Mean Sentiment and Mean Manually Coded Scores per Questionaire Section



Analysis of Summarization

In order to answer the second sub-question on whether text summarization can improve the analysis of qualitative student feedback, the PEGASUS summarization was evaluated. Therefore, descriptive statistics were first plotted. Overall, there were 138 comments. As can be seen in Table 4, the mean summary rating was M=1.82 with a standard deviation of SD=.21. Moreover, the mean comment length in words was M=43.73 with a standard deviation of SD=39.48, in contrast to the average summary which contained M=13.54 words on average (SD=8.05). Therefore, the summarization reduced the comment length by a third on average.

Table 3.

Descriptive Statistics for the PEGASUS Model

Variable	*M*	*SD*	Minimum	Maximum
1: Manual Summary Rating	1.82	.21	1.00	5.00
2: Comment Length	43.73	39.48	3	283
3: Summary Length	13.54	8.05	3	50

Discussion

The goal of this research was to provide an understanding of how text mining can help in the process of evaluating student feedback. This objective was reached by answering the research question "*How can text mining assist educational institutions in the process of analysis of student evaluations of teaching?".* To make this question more tangible, three sub-questions were drafted which addressed state-of-the-art text mining methods for student feedback. Firstly, a sentiment analysis and secondly a text summarization was implemented. Third, a literature review was conducted to map the current state of literature on this topic. In this chapter, the evaluation of these will be discussed Additionally, the limitations of this study will be examined, as well as the potential for future research for text mining student feedback.

Sentiment Analysis

The first sub-question of this research regarded how a sentiment analysis could assist in the process of analysing student feedback. The data here indicated that the scoring of a sentiment analysis on written student feedback is significantly associated with corresponding Likert-type data. This was confirmed by correlating the mean Likert-type scores per section with mean sentiment scores per section. What these results generally indicate is that a sentiment analysis quantitative of textual feedback might be a sufficient substitute for quantitative measurements in student feedback. In other words, Likert-type questions might be obsolete and hence replaceable by asking for open-ended questions only, by implementing a sentiment analysis. Furthermore, it is expected that such a sentiment analysis system could be applied by practitioners in real-life situations with high efficacy. Overall, these findings are in line with the previous research by Neumann & Linzmayer (2021), who predicted a five-star measurement with the use of sentiment analysis on textual feedback. In contrast to Neumann & Linzmayer (2021) however, this study did not gather distinct feedback only dedicated to this research but instead analysed the natural responses to formal student feedback that students do get asked every semester.

Another novelty of this approach was that it was one of the first to perform sentiment analysis on previously topic-categorized comment data. The results are therefore able to show sentiment analysis on pre-categorized comment data can predict corresponding mean scores. Unlike Gottipati et al. (2017) the sentiment analysis tool did not extract topics from the comments but used pre-structured data. For universities, this reduces the need for a further technical categorization solution. Additionally, this also gives students the opportunity to provide their feedback in a structured way, which has been associated with the production of deeper feedback compared to free-text-only formats (Hoon et al., 2014).

In addition, one aim of this study was to understand the ability of a sentiment analysis to accurately capture the emotions of students' comments in feedback surveys. Therefore, the sentiment scores of one of the student feedback reports were correlated with the ratings made by a human assessor. The results then displayed a strong relationship between these two variables. These findings indicate that sentiment analysis is an accurate tool in measuring students' emotions as expressed in textual student feedback.

Text Summarization

The second text mining method that this study explored was text summarization. Here the results showed that the summarizer performed poorly to moderately, consequently not being a useful tool for accurately capturing the content of student comments. An explanation for these findings might be the fact that the PEGASUS summarization model that was used has only been trained on online news site content. Because of that, a fine-tuned model might overall be able to produce better results. Another reason for the model's poor performance could be that students on average only submitted a small body of text for feedback. Summarization however is generally used to extract the main idea of lengthy text documents (Gaikwad et al., 2014). Providing the system with too little information could therefore have resulted in its poor performance.

Systematic Literature Review

The third sub question of this research regarded what methodologies are used for the analysis of student feedback and what additional techniques apart from sentiment analysis and text summarization could be of use. To answer this question, ten research articles on text mining of student feedback were reviewed. It became apparent that the most well-researched approaches are both sentiment analysis and clustering, with categorization being slightly less well-researched. Additionally, none of the literature regards text summarization, consequently indicating that this method currently holds great potential for further exploration. At last, no additional text mining method was found, apart from sentiment analysis, clustering and categorization. In general, these findings mimic the results of Ferreira-Mello et al. (2019), who investigated the frequency of text mining applications in education and found text summarization to be among the least applied ones.

Limitations and Future Research

This study had several limitations. Firstly, for the sentiment analysis there were technical limitations in regards to validity, as the sentiment model had been pretrained on product reviews, and was consequently not attuned to educational text data. Secondly, small sample sizes also create threats to the validity of a study. Since the first part of this study was conducted on a comparison of 35 scores, the generalizability of this data is only limited. Additionally, the comparison was done only on the means of the original data, which was composed of more than 400 comments. Thirdly, there are some limitations in regards to the reliability of human assessment in both the sentiment analysis and text summarization, due to the fact that there was only one assessor. One final limitation also is that this research mainly focused on only two text mining techniques.

Taking all these limitations into account, future research should therefore be focused on extending this work in three ways. Firstly, by also applying other text mining methods, such as categorization and clustering algorithms, as well as the less well-researched method of text summarization. Secondly, the use of larger data sets might further allow the finetuning of pretrained models for the tasks of analysing student feedback. Zhang et al. (2020) for example assert that a PEGASUS model can become highly capable by training it on as little as 1000 examples. The third direction for future research is the development of a conceptual framework for text mining in educational feedback. More work might therefore be done in providing researchers with conceptual tools in the analysis of what techniques are needed and how they can solve the problems of analysing large qualitative data sets. On top of that, newly developed prototypes might then be tested with important stakeholders in the actual university environment.

On a final note, previous research showed how students' emotional states hold predictive power over quantitative ratings of a course (Wachtel, 1998). Similarly, this study indicated the predictive power of sentiment analysis scores over corresponding numerical course ratings. The implication is therefore that it is not clear whether quantitative measurements can indeed measure emotional states. That would mean that emotions are confounding quantitative scores and scores of a comment sentiment analysis. Future research may therefore clarify the role of emotions in both quantitative measures, as well as in sentiment analysis.

Conclusion

This research studied the usage of novel text mining techniques in analysing textual student feedback. For the limitation of the thesis' time frame, only two of these methods were chosen and applied on previously gathered student feedback from the University of Twente. The criterion for the choice of these methods was that, in contrast to other methods e.g. clustering, they showed larger research gap for text summarization and the comparison of sentiment scores with Likert-type scores. Next to the application of these two methods, a literature study was conducted to map the applicability of other techniques. Therefore, this research consisted of three major parts.

Firstly, a sentiment analysis was conducted on all comments and compared with corresponding quantitative data as well as the polarity ratings of a manual assessor. The results then indicated that sentiment analysis scores are strongly associated with corresponding quantitative measurements, as well as manual polarity assessments. Additionally, the results showed that a sentiment analysis on pre-structured data is effective. The implication of that is that providing students with a structure for their text feedback is advantageous for its later analysis. Next to the sentiment analysis, a text summarization technique was employed on students' comments. In contrast to the sentiment analysis, this did not produce meaningful information for further evaluation. These results were accounted for by the short length of the comments, which opposed the design of the summarization system on lengthy text documents. Third, a literature review was conducted to map the currently existing text mining methodologies in analysing textual student feedback. This systematic review revealed that most research incorporates either sentiment analysis, clustering, or categorization. No other methodology was found. Furthermore, this study pointed out areas for future research, mainly pertaining to increasing the validity of these results. In conclusion, the field of analysing student feedback holds great potential for the development of text mining applications.

Acknowledgements

I would like to thank my first supervisor Dr Gayane Sedrakyan and my second supervisor Dr Maryam Amir Haeri for their supervision during this thesis. I am especially grateful to Dr Sedrakyan who not only gave me the opportunity to delve into a completely new field of research but also challenged me with the freedom to try different text mining techniques. Since this project did not happen within the streamlined graduation process of the psychology program, I can hardly understate the support that Dr Serakyan gave me in making this project possible. Moreover, Dr Serakyan always took the time to clarify any question I had and pushed me to try out a variety of novel methodologies. Finally, I would like to thank my parents, friends and especially my girlfriend Louisa, who supported me throughout the process of writing this thesis.

References

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Aung, K. Z., & Myo, N. N. (2017). Sentiment analysis of students' comment using lexicon based approach. *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. https://doi.org/10.1109/icis.2017.7959985

Beke, A., & Szaszák, G. (2016). Automatic Summarization of Highly Spontaneous Speech. *Speech and Computer*, 140–147. https://doi.org/10.1007/978-3-319-43958-7_16

Bhaduri, S., Soledad, M., Roy, T., Murzi, H., & Knott, T. (2021). A Semester Like No Other: Use of Natural Language Processing for Novice-Led Analysis on End-of-Semester Responses on Students' Experience of Changing Learning Environments Due to COVID-19. *2021 ASEE Virtual Annual Conference Content Access Proceedings*. https://doi.org/10.18260/1-2--36609

Brockx, B., van Roy, K., & Mortelmans, D. (2012). The Student as a Commentator: Students' Comments in Student Evaluations of Teaching. *Procedia - Social and Behavioral Sciences*, *69*, 1122–1133. https://doi.org/10.1016/j.sbspro.2012.12.042

Cen, W., Gao, Z., Xu, R., Wu, B., Zheng, L., Zhao, W., Xiao, L., & He, X. (2021). Extraction Method for Constructive Proposals based on Online Comments. *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. https://doi.org/10.1109/dasc-picom-cbdcom-cyberscitech52372.2021.00147

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *WIREs Data Mining and Knowledge Discovery, 9*(6). https://doi.org/10.1002/widm.1332

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications*, *85*(17), 42–45. https://doi.org/10.5120/14937-3507

Google. (n.d.). *Pegasus*. HuggingFace. Retrieved June 7, 2022, from https://huggingface.co/google/pegasus-xsum

Gottipati, S., Shankararaman, V., & Gan, S. (2017). A conceptual framework for analyzing students' feedback. *2017 IEEE Frontiers in Education Conference (FIE)*. https://doi.org/10.1109/fie.2017.8190703

Gottipati, S., Shankararaman, V., & Lin, J. R. (2018). Text analytics approach to extract course improvement suggestions from students' feedback. *Research and Practice in Technology Enhanced Learning, 13*(1). https://doi.org/10.1186/s41039-018-0073-0

Gronberg, N., Knutas, A., Hynninen, T., & Hujala, M. (2021). Palaute: An Online Text Mining Tool for Analyzing Written Student Course Feedback. *IEEE Access*, *9*, 134518–134529. https://doi.org/10.1109/access.2021.3116425

Harvey, L. (2003). Student Feedback. *Quality in Higher Education*, *9*(1), 3–20. https://doi.org/10.1080/13538320308164

Hoon, A., Oliver, E., Szpakowska, K., & Newton, P. (2014). Use of the 'Stop, Start, Continue' method is associated with the production of constructive qualitative feedback by students in higher education. *Assessment & Evaluation in Higher Education*, *40*(5), 755–767. https://doi.org/10.1080/02602938.2014.956282

Hynninen, T., Knutas, A., & Hujala, M. (2020). Sentiment analysis of open-ended student feedback. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. https://doi.org/10.23919/mipro48935.2020.9245345

Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Applied Sciences, 11*(9), 3986. https://doi.org/10.3390/app11093986

Katz, A., Norris, M., Alsharif, A. M., Klopfer, M. D., Knight, D. B., & Grohs, J. R. (2021, July). *Using Natural Language Processing to Facilitate Student Feedback Analysis*. 2021 ASEE Virtual Annual Conference. https://peer.asee.org/using-natural-language-processing-to-facilitate-student-feedback-analysis

Kember, D., Leung, D. Y. P., & Kwan, K. P. (2002). Does the Use of Student Feedback Questionnaires Improve the Overall Quality of Teaching? *Assessment & Evaluation in Higher Education*, *27*(5), 411–425. https://doi.org/10.1080/0260293022000009294

Loderer, K., Pekrun, R., & Lester, J. C. (2020). Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. *Learning and Instruction*, *70*, 101162. https://doi.org/10.1016/j.learninstruc.2018.08.002

Luo, W., Liu, F., Liu, Z., & Litman, D. (2016). Automatic Summarization of Student Course Feedback. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://doi.org/10.18653/v1/n16-1010

Mandouit, L. (2018). Using student feedback to improve teaching. *Educational Action Research*, *26*(5), 755–769. https://doi.org/10.1080/09650792.2018.1426470

Mathew, L., & Bindu, V. R. (2020). A Review of Natural Language Processing Techniques for Sentiment Analysis using Pre-trained Models. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. https://doi.org/10.1109/iccmc48092.2020.iccmc-00064

Mega, C., Ronconi, L., & de Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology*, *106*(1), 121–131. https://doi.org/10.1037/a0033546

Nasim, Z., Rajput, Q., & Haider, S. (2017). Sentiment analysis of student feedback using machine learning and lexicon based approaches. *2017 International Conference on*

Research and Innovation in Information Systems (ICRIIS). https://doi.org/10.1109/icriis.2017.8002475

Nawaz, R., Sun, Q., Shardlow, M., Kontonatsios, G., Aljohani, N. R., Visvizi, A., & Hassan, S. U. (2022). Leveraging AI and Machine Learning for National Student Survey: Actionable Insights from Textual Feedback to Enhance Quality of Teaching and Learning in UK's Higher Education. *Applied Sciences, 12*(1), 514. https://doi.org/10.3390/app12010514

Neumann, M., & Linzmayer, R. (2021). Capturing Student Feedback and Emotions in Large Computing Courses: A Sentiment Analysis Approach. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. https://doi.org/10.1145/3408877.3432403

Nitin, G. I., Swapna, G., & Shankararaman, V. (2015). Analyzing educational comments for topics and sentiments: A text analytics approach. *2015 IEEE Frontiers in Education Conference (FIE)*. https://doi.org/10.1109/fie.2015.7344296

NLP Town. (n.d.). *Bert-base-multilingual-uncased-sentiment*. Hugging Face. Retrieved April 1, 2022, from https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Onan, A. (2019). Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education*, *28*(1), 117–138. https://doi.org/10.1002/cae.22179

Palmer, S., & Campbell, M. (2013). *Practically and productively analysing Course Experience Questionnaire student comment data*. Australasian Association for Engineering Education. Conference, Gold Coast, Queensland. https://dro.deakin.edu.au/view/DU:30060776

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review, 18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education*, *30*(4), 387–415. https://doi.org/10.1080/02602930500099193

Rani, S., & Kumar, P. (2017). A Sentiment Analysis System to Improve Teaching and Learning. *Computer*, *50*(5), 36–43. https://doi.org/10.1109/mc.2017.133

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1355

Rowley, J. (2003). Designing student feedback questionnaires. *Quality Assurance in Education*, *11*(3), 142–149. https://doi.org/10.1108/09684880310488454

Sadriu, S., Nuci, K. P., Imran, A. S., Uddin, I., & Sajjad, M. (2022). An Automated Approach for Analysing Students Feedback Using Sentiment Analysis Techniques. *Pattern Recognition and Artificial Intelligence*, 228–239. https://doi.org/10.1007/978-3-031-04112-9_17

Santhanam, E., Lynch, B., & Jones, J. (2018). Making sense of student feedback using text analysis – adapting and expanding a common lexicon. *Quality Assurance in Education*, *26*(1), 60–69. https://doi.org/10.1108/qae-11-2016-0062

Scott, G. (2021). Accessing the student voice. *Analysing Student Feedback in Higher Education*, 149–163. https://doi.org/10.4324/9781003138785-13

Shah, M., & Pabel, A. (2019). Making the student voice count: using qualitative student feedback to enhance the student experience. *Journal of Applied Research in Higher Education*, *12*(2), 194–209. https://doi.org/10.1108/jarhe-02-2019-0030

Toçoğlu, M. A., & Onan, A. (2020). Sentiment Analysis on Students' Evaluation of Higher Educational Institutions. *Advances in Intelligent Systems and Computing*, 1693–1700. https://doi.org/10.1007/978-3-030-51156-2_197

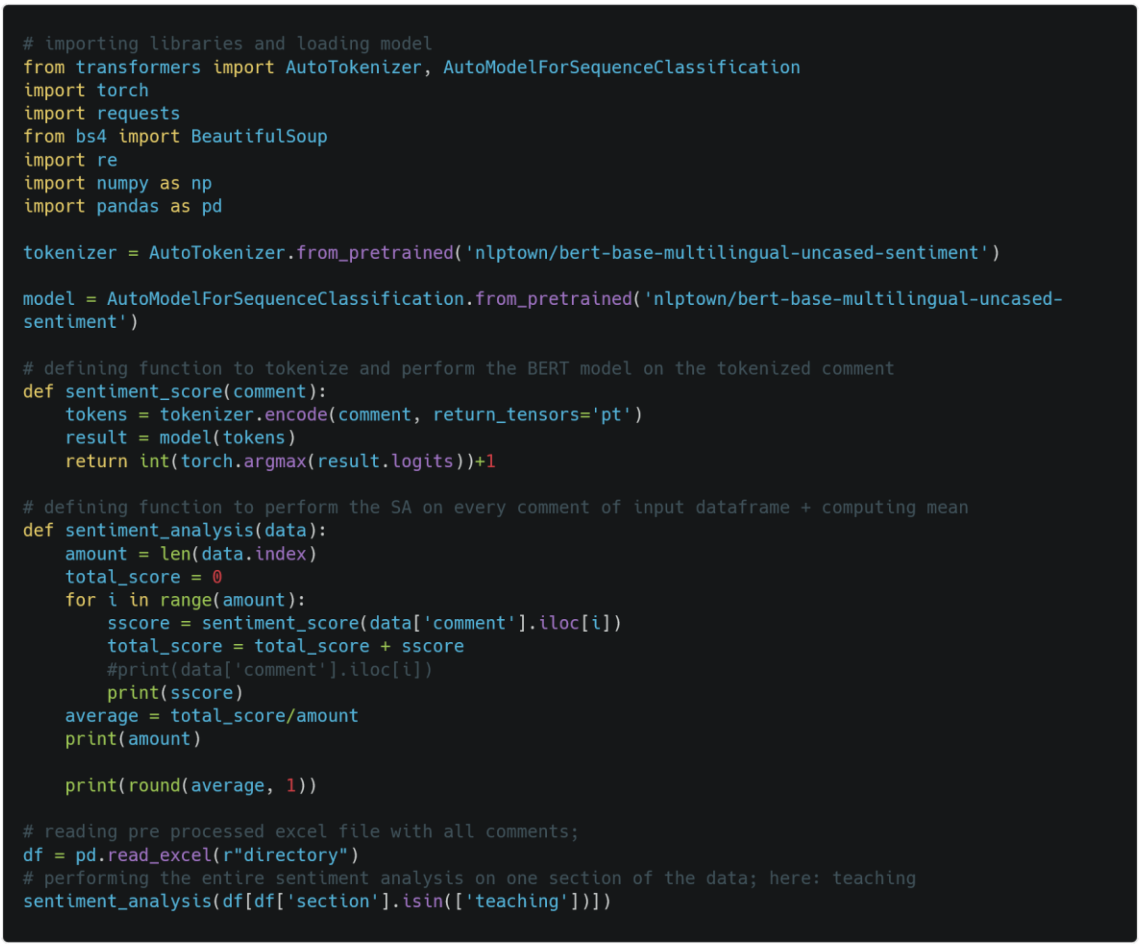
Tyng, C. M., Amin, H. U., Saad, M. N. M., & Malik, A. S. (2017). The Influences of Emotion on Learning and Memory. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.01454

Ulfa, S., Bringula, R., Kurniawan, C., & Fadhli, M. (2020). Student Feedback on Online Learning by Using Sentiment Analysis: A Literature Review. *2020 6th International Conference on Education and Technology (ICET)*. https://doi.org/10.1109/icet51153.2020.9276578

Wachtel, H. K. (1998). Student Evaluation of College Teaching Effectiveness: a brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191–212. https://doi.org/10.1080/0260293980230207

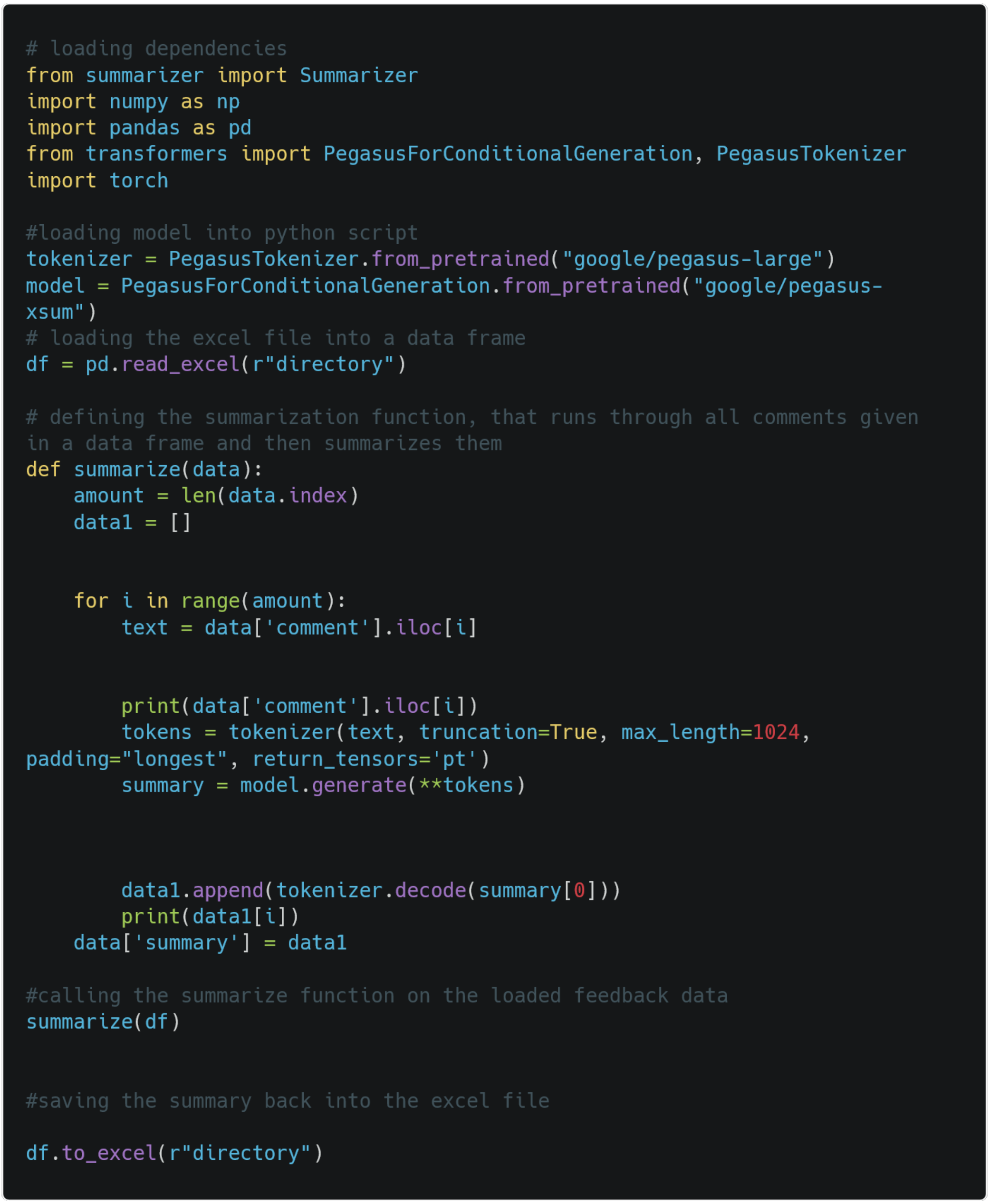
Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. International Conference on Machine Learning, Vienna, Austria. https://doi.org/10.48550/arXiv.1912.08777

Appendix A: Python Script for Sentiment Analysis



```
# importing libraries and loading model
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch
import requests
from bs4 import BeautifulSoup
import re
import numpy as np
import pandas as pd

tokenizer = AutoTokenizer.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')

model = AutoModelForSequenceClassification.from_pretrained('nlptown/bert-base-multilingual-uncased-
sentiment')

# defining function to tokenize and perform the BERT model on the tokenized comment
def sentiment_score(comment):
    tokens = tokenizer.encode(comment, return_tensors='pt')
    result = model(tokens)
    return int(torch.argmax(result.logits))+1

# defining function to perform the SA on every comment of input dataframe + computing mean
def sentiment_analysis(data):
    amount = len(data.index)
    total_score = 0
    for i in range(amount):
        sscore = sentiment_score(data['comment'].iloc[i])
        total_score = total_score + sscore
        #print(data['comment'].iloc[i])
        print(sscore)
    average = total_score/amount
    print(amount)

    print(round(average, 1))

# reading pre processed excel file with all comments;
df = pd.read_excel(r"directory")
# performing the entire sentiment analysis on one section of the data; here: teaching
sentiment_analysis(df[df['section'].isin(['teaching'])])
```

Appendix B: Python Script for Text Summarization



```
# loading dependencies
from summarizer import Summarizer
import numpy as np
import pandas as pd
from transformers import PegasusForConditionalGeneration, PegasusTokenizer
import torch

#loading model into python script
tokenizer = PegasusTokenizer.from_pretrained("google/pegasus-large")
model = PegasusForConditionalGeneration.from_pretrained("google/pegasus-
xsum")
# loading the excel file into a data frame
df = pd.read_excel(r"directory")

# defining the summarization function, that runs through all comments given
in a data frame and then summarizes them
def summarize(data):
    amount = len(data.index)
    data1 = []


    for i in range(amount):
        text = data['comment'].iloc[i]


        print(data['comment'].iloc[i])
        tokens = tokenizer(text, truncation=True, max_length=1024,
padding="longest", return_tensors='pt')
        summary = model.generate(**tokens)



        data1.append(tokenizer.decode(summary[0]))
        print(data1[i])
    data['summary'] = data1

#calling the summarize function on the loaded feedback data
summarize(df)


#saving the summary back into the excel file

df.to_excel(r"directory")
```

Appendix C: Final Literature List for Systematic Review

Table C1.

Study Characteristics and Methods used

Author	Method	Location	Displaying results	Evaluation	Description
Bhaduri et al. (2021)	Information retrieval, Sentiment analysis,	United States	Emoji next to a frequently used term	/	Trying a text mining approach to evaluate course feedback in response to COVID-19
Gottipati et al. (2017)	Clustering, Categorization, Sentiment Analysis	Singapore	Topic-based sentiment bar chart	?	Providing a framework for the automation of qualitative feedback analysis
Gottipati et al. (2018)	Categorization	Singapore	Word-Cloud, table with search bar for comments	F-Scores	Detecting explicit and implicit suggestions; Comparing different Methods
Gronberg et al. (2021)	Sentiment analysis, Clustering	Finnland	Summary for clusters with bar charts on type of emotion, key words and comments included in the cluster	?	Developing tool for automated analysis of qualitative data in student feedback.

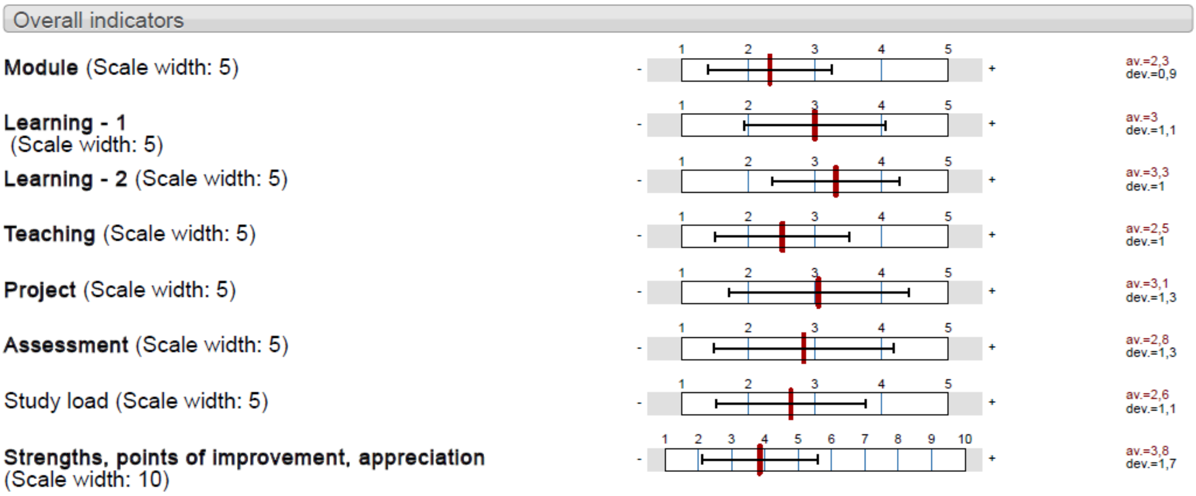
Hynninen et al. (2020)	Sentiment Analysis	Finnland	Word-Cloud	/	Analysing the type of emotions in student feedback
Katz et al. (2021)	Clustering	United States	/	Automatically identified topics were compared against manually coded topics.	Developing and evaluating a clustering assissted way of coding textual student feedback
Nawaz et al. (2022)	Categorization	United Kingdom	/	Comparison of improvement in end-of year evaluations after implementing intervention	Extracting actionable feedback to design a teaching intervention
Neumann & Linzmayer (2021)	Sentiment Analysis	United States	/	Mean absolute difference between manual assessment and sentiment scores	Measuring emotions in student feedback with sentiment analysis
Nitin et al. (2015)	Clustering, Categorization, Sentiment Analysis	Singapore	Topic-based sentiment bar chart	?	Developing a feedback mining system

Shah & Pabel (2019)	Clustering	Australia	Visual concept map with terms clustered	/	Gaining insights into the experience of online and on-campus students

Appendix D: Example of Likert-Type Scores in Feedback Sheet

Figure D1.

Example of Likert-Type Scores in Student Feedback Sheet



Overall indicators

Module (Scale width: 5)	av.=2,3	dev.=0,9
Learning - 1 (Scale width: 5)	av.=3	dev.=1,1
Learning - 2 (Scale width: 5)	av.=3,3	dev.=1
Teaching (Scale width: 5)	av.=2,5	dev.=1
Project (Scale width: 5)	av.=3,1	dev.=1,3
Assessment (Scale width: 5)	av.=2,8	dev.=1,3
Study load (Scale width: 5)	av.=2,6	dev.=1,1
Strengths, points of improvement, appreciation (Scale width: 10)	av.=3,8	dev.=1,7

Appendix E: Score Comparison per Section

Table D1.

Comparison of Mean Scores per Questionnaire-Section

Section		*Mean Sentiment Score*	*Mean Quantitative Score*
Module	Mean	2.52	3.22
	SD	.48	.57

Learning 1	Mean	3.04	3.56
	SD	.35	.38
Learning 2	Mean	3.60	3.70
	SD	.45	.27
Teaching	Mean	2.80	3.22
	SD	.30	.47
Project	Mean	3.52	3.60
	SD	.15	.36
Assessment	Mean	2.54	3.38
	SD	.64	.42
Study Load	Mean	2.80	3.26
	SD	.16	.59