Studies in Educational Evaluation 83 (2024) 101406

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Sentiment analysis of student feedback: A comparative study employing lexicon and machine learning techniques

Charalampos Dervenis [a,*], Giannis Kanakis [b], Panos Fitsilis [c,d]

[a] Department of Business Administration, Schools of Economics and Business Administration, University of Thessaly, Larisa, Greece
[b] Computer Science Department, Hellenic Open University, School of Science & Technology, Patras, Greece
[c] Computer Science Department, Hellenic Open University, Patras, Greece
[d] Department of Business Administration, University of Thessaly, Larisa, Greece, Hellenic Open University, Patras, Greece

ARTICLE INFO

Keywords:
Student feedback
Teaching evaluation
Educational data
Sentiment analysis
Lexicon based approach
Machine learning approach

ABSTRACT

One of the most pressing concerns in contemporary education examines the integration of big data and artificial intelligence methodologies to enhance the educational learning outcome. Towards that purpose, it is imperative to leverage the unstructured data originating from student feedback, particularly in the form of comments to open-ended questions aiming to extract emotions and opinions conveyed within their messages. Our research goal is to ascertain the most efficient approach to tackle this difficult task by conducting a comparison of sentiment analysis methods, including Machine Learning (ML) and Lexicon based models. Both lexicon based and machine learning approaches were implemented using an open source data mining platform while also utilizing student comments submitted at the end of academic semesters. Our study reveals a promising approach that effectively addresses the issue at hand, particularly within the domain of educational data. Additionally, it emphasizes the key aspects that led to the selection of this approach effectively highlighting the weaknesses and strengths inherent in each method

1. Introduction

The amount of data generated within the four-year period (2020–2023) surpasses the cumulative data produced throughout the entirety of human history until the year 2023 (Data growth worldwide, 2010–2025 & Statista,10–2025; Statista, 2023). This notable surge in data volume can be attributed to advances in technology, the extensive adoption of the internet, the prevalence of social media platforms, the proliferation of sensors and various other sources that perpetually generate overwhelming amounts of data. The field of education is no exception to the above trend. A growing number of educational institutions are embracing internet based learning platforms, such as Learning Management Systems (LMS) and other web applications (Wei, 2023). These platforms and applications serve as digital frameworks for managing and delivering educational content, thereby facilitating online learning, collaboration, and communication between educators and students. Within these learning ecosystems, students are encouraged to provide written feedback, allowing them to address specific issues or concerns related to the learning process.

These comments and feedback can be gathered either in real time, during the teaching process, or through the forums that are often provided by the aforementioned learning platforms. Furthermore, the practice of soliciting evaluations from students at the end of each semester has become a standard procedure in higher education institutions. Its primary objective is to evaluate and improve the quality of teaching. These evaluations aid instructors in refining their teaching techniques and gaining a deeper understanding of students' perspectives. An educator can receive feedback on their teaching, identify students in need of additional support, predict student performance, and group the information to uncover learning patterns. Furthermore, common remarks made by students can be highlighted, and improvements in the educational process can be sought. Based on the discussion above, it becomes evident that an effective way for an educator to enhance their teaching methodologies is to obtain timely and candid feedback from their students. However, the manual processing of responses to open-ended questions can be time consuming and challenging. Thus, in the era of blended and online teaching approaches, characterized by the lack of emotional communication, it is becoming

* Correspondence to: Department of Business Administration, School of Economics and Business, UNiversity of Thessaly, Gaiopolis Campus, Larissa ring road, GR41500 Larissa, Greece.
E-mail address: dervenisch@gmail.com (C. Dervenis).

https://doi.org/10.1016/j.stueduc.2024.101406
Received 22 May 2024; Received in revised form 16 September 2024; Accepted 20 September 2024
Available online 2 October 2024

imperative to adopt methods and foster ecosystems that facilitate the extraction of valuable insights from this kind of textual data (Dervenis et al., 2022). In such cases, sentiment analysis techniques can be applied to isolate the ideas and observations contained within the messages in question.

This paper investigates the application of sentiment analysis techniques in evaluating comments provided by students in response to open-ended questions, which are typically submitted at the end of each academic semester. Numerous studies (Ganganwar & Rajalakshmi, 2023; Hu & Liu, 2004; Misuraca et al., 2021; Wook et al., 2020), have proposed a methodology for polarity categorization, employing parts of speech (POS) and differentiating between positive and negative words. Additionally, a multitude of methods utilizing machine learning algorithms to address the same issue have been put forward (Pang et al., 2002; Salmony et al., 2023; Sebastiani, 2002). Gamallo and Garcia (2014), as well as Gautam and Yadav (2014), proposed a Naive Bayes classifier while Joachims (1998), introduced a method based on Support Vector Machines (SVM). Consequently, an opportunity arises for performing a comparative analysis among these primary approaches. Through this research, we conducted a comparative analysis of Machine Learning (ML) and Lexicon-based models in order to determine the most effective and efficient method for sentiment analysis of student responses in course quality questionnaires. The primary questions driving our research are as follows:

RQ1: Is there a competitive advantage in terms of classification accuracy, multilingual support and context sensitivity of ML approaches over Lexicon based approaches?
RQ2: How does the inclusion of neutral sentiment class affect the performance of classifiers in each approach?

Through our investigation, we aim to provide insights and recommendations regarding the choice of approach for sentiment analysis in this specific context. In the next sections, we introduce the field of Sentiment Analysis in Section 2, where we also provide an overview of the methods employed in this area. In Section 3 we analyze the methodology employed to obtain our research outcomes. In Section 4, we present our implementation, aiming to emphasize our methodology. Finally, in Section 5, we use tables to concisely present the obtained results and relevant information. We also present both our limitations as well as closing remarks, encapsulating the key findings and implications of our study.

2. Background

2.1. Sentiment analysis

Sentiment Analysis (SA) can be referred to by various names in the literature. It may be mentioned as subjectivity analysis, opinion mining, or appraisal extraction (Misuraca et al., 2021). According to Liu (2012), "Sentiment analysis is the field of study that analyzes people's opinions, emotions, evaluations, appraisals, and attitudes towards entities such as products, services, organizations, individuals, topics, events, and their attributes.". Essentially, SA refers to the process of taking a text or text snippet as input and extracting valuable information related to the expressed sentiments within that text. This process is typically performed on a collection of texts referred to as a corpus. Sentiment analysis plays a significant role in many applications of business intelligence, such as credit rating assessment or organizational reputation evaluation. It is essential for the development of analytics that aim to assess the limitations of offered products and services in order to improve their quality. Furthermore, SA is perhaps one of the most widespread use cases of descriptive analysis. As a research field, sentiment analysis can be considered part of computational linguistics and natural language processing (NLP), which involves the interaction between computers and human (natural) language (Mejova, 2009; Saraswat et al., 2020). It employs techniques and methods from machine learning, natural language processing and statistics, with the goal of classifying text into sentiment classes. The sentiment polarity is one of the desired properties of the text being processed. It constitutes the main methodological characteristic of sentiment analysis and is commonly categorized into three classes, Positive, Neutral and Negative (Chakraverty & Saraswat, 2017; Garg, 2020; Liu, 2012; Wilson et al., 2009). However, this does not imply that binary classification (positive, negative) or multiclass classification with more categories, (i.e., Strongly Positive, Positive, Neutral, Negative, Strongly Negative) cannot be applied. The number of classes in sentiment analysis can vary depending on the specific needs and context of the analysis. Essentially, the number of classes used for text categorization is dictated by the specific problem at hand. Furthermore, besides polarity identification, there are more sophisticated techniques that recognize emotional states in the text, such as "joy, " "sadness," or "anger." Finally, it is important to distinguish between polarity and intensity (strength) with which the sentiment is expressed in the text. For example, one might have a positive opinion about a professor but not hold a strong opinion since they have not attended their class for a significant period of time.

Sentiment Analysis can play a pivotal role in various facets of the learning process. Not only does it allow educators to receive feedback on their instructional methods, but it can also facilitate the identification of students requiring extra support (Altrabsheh et al., 2013; Dervenis et al., 2024; Poulos & Mahony, 2008; Wen et al., 2014). The state of the art in sentiment analysis distinguishes three dominant approaches aimed at addressing this concern. These approaches are briefly yet concisely presented in the following section.

2.2. Algorithms and techniques used in sentiment analysis

2.2.1. Machine learning-based approaches

There are several types of machine learning algorithms applied to classification problems, including sentiment analysis. In this research, we focus on random forest (RF) as well as on the artificial neural networks (ANNs). Random forests are a modern machine learning method that belongs to the category of supervised learning algorithms and performs exceptionally well in classification problems. The first algorithm, as implied by its name, consists of an ensemble of decision trees. Each individual decision tree in the random forest makes a class prediction, and the class with the most votes becomes the final prediction of the model. The fundamental idea behind the method is that a large number of uncorrelated models functioning as a committee will outperform any of the individual constituent models. The low correlation among the models is key, as the uncorrelated models can produce predictions that are more accurate than any of the individual predictions (Dervenis, et al., 2022). This is because the trees compensate for each other's errors through their overlapping nature. Furthermore, unlike individual decision trees, the decision trees belonging to a random forest are constructed by randomly selecting predictor variables based on which the splitting of nodes in the tree occurs. These characteristics, combined with the absence of pruning, result in lower correlation among the trees and, therefore, greater diversity among them. Random forests, and subsequently decision trees, belong to the category of nonparametric machine learning algorithms.

Artificial neural networks are a form of machine learning algorithm that seeks to emulate, to the extent feasible, the operation of the human brain in order to solve complex problems. The mathematical model underpinning this particular method strives to replicate the structure of the biological neural network, thereby acquiring characteristics that enable parallel data processing and the execution of tasks that are challenging to describe, such as pattern recognition. ANNs represent a relatively new area of research that began its development on an international level a few decades ago, around 1980. Initially, McCulloch and Pitts (1943), presented a model of a neural network based on neurons and their connections. This model forms the basis of modern

artificial neural networks, which have the perceptron as their basic structural unit and "learn" through processes that modify the weights of their connections (synaptic weights). After the learning process, the network can undergo the recall process and compute its output for a given input vector. In other words, it applies the learned model. The use of these techniques requires a large volume of data, and most of them rely on correctly labeled data during the training process. In contrast, lexicon based techniques can avoid this limitation.

2.2.2. Lexicon-based approaches

The lexicon based approach utilizes a lexicon that associates words with corresponding sentiment weights, capturing the meaning attributed to each word. Depending on the implementation, weights can take two values (positive/negative sentiment) or belong to a range of values. The polarity values of each word in the text are transmitted to an algorithm to generate the final sentiment score, which represents the sentiment conveyed by the text. We can distinguish two main approaches in the application of Lexicon-Based methods based on the way sentiment bearing words are identified during the creation of the lexicon (Kausar et al., 2019). The first approach is the Lexicon-based approach, in which lexicons can be created either manually or automatically using seed words, typically adjectives, as indicators of the semantic orientation conveyed by the text. These seed words are then expanded with their synonyms to enrich the initial lexicon as much as possible. The implementation of this technique can be further expedited by utilizing a preexisting lexicon from a variety of resources available for sharing, primarily in the English language. Examples of such lexicons include NRC Emotion Lexicon, MPQA, WordNet, SenticNet, SentiWordNet, WordNet-Affect, and others (Hardeniya & Borikar, 2016). Furthermore, we have the corpus based approach that aids in identifying words that carry sentiment by utilizing heuristic rules to comprehend the meaning of a word based on its context, ultimately assigning it to the corresponding sentiment category, which involves calculating the polarity of a word based on its co-occurrence frequency with another word of known polarity.

2.2.3. Hybrid approaches

Hybrid approaches combine the two aforementioned techniques to improve the overall performance. A powerful example of this technique is the work of Xia et al. (2011), where they propose a hybrid technique for sentiment analysis. Initially, they use lexicon-based analysis and NLP techniques for word categorization based on their part of speech (POS tagging). Then, they apply three different machine learning algorithms to estimate the sentiment expressed in the text.

3. Methodology and sentiment analysis approach

In this section, we provide a comprehensive overview of the stages involved in the process we followed for each SA approach. The visual depiction of these stages is presented in Fig. 1. Specifically, in the context of machine learning algorithms, there is a requirement to incorporate additional steps pertaining to feature selection and model training. Each of these stages are described in detail in the following paragraphs. We have chosen to depict the overall methodology using a circular pattern instead of a linear one to highlight the iterative nature of the procedure, especially in terms of model fine tuning. This circular representation signifies that the different stages of the methodology are not necessarily one time tasks, but rather an ongoing cycle of refinement and optimization.

3.1. Data acquisition

According to our University's regulations, and as part of improving the operations of the academic unit, it is essential to evaluate each course and instructor based on student feedback, as well as to thoroughly assess the results of this evaluation. This assessment is conducted

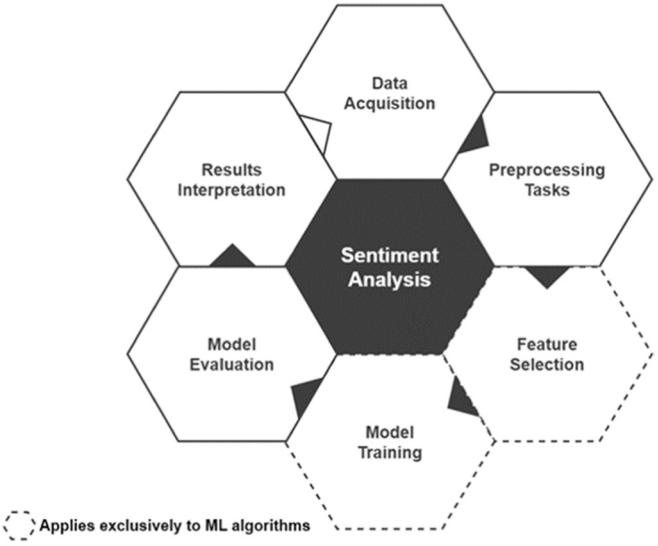

Fig. 1. The overall procedure followed.

through an online questionnaire, which has been carefully designed by the University's Quality Assurance Unit. The anonymity of the participating students is fully guaranteed throughout the process.

Specifically, at the end of each semester, students are asked to anonymously complete an online questionnaire for every course and instructor. A link to the questionnaire, aligned with the guidelines of the National Higher Education Assessment Authority of Greece, is sent via email to students by the Quality Assurance Unit of the University of Thessaly. The questionnaire consists of both closed and open-ended questions. The data collected from these responses is stored online and is accessible to authorized administrators. This dataset, depending on the system used for the analysis purposes, can be imported in various file formats such as Excel (.xlsx), comma-separated (.csv), native tab-delimited (.tab) files and others.

In our case, the data collection was conducted at the end of the semester, specifically in June 2023, with strict adherence to the General Data Protection Regulation (GDPR) which ensured the complete anonymity of students throughout the process. The data used in the present study, consists of text generated by students in response to the following questions:

1) Please provide general comments and suggestions that you would like to convey to your professor.
2) Please provide general comments and suggestions you would like to make regarding the functioning of the course.

After excluding empty responses, we obtained a total of 1420 responses from the initial pool of 1680 students. Out of the 1420 responses, 652 were responses to the first question, while the remaining were responses to the second question. In order to evaluate the performance of the methods we applied, both the lexicon-based and classification algorithms, we needed to have actual sentiment labels for our data available. This allowed us to calculate various performance metrics through the corresponding confusion matrices. These actual sentiment labels, referred to as the "true sentiment" (e.g., positive, negative, neutral), must be predefined either through manual annotation from "evaluators" or reliable sources. At the same time, each method (lexicon or classification algorithms) must generate its own predictions for the sentiment of each data sample (e.g., positive, negative, neutral) to assess its performance. Next, for the calculation of metrics regarding the performance of each sentiment analysis method, the predicted label is taken as the one generated by the sentiment analysis algorithm, while the real label is the one pre-determined by the evaluators.

For this purpose, the obtained responses were randomly arranged and made available to two evaluators whose task was to manually classify the text of each response into one of three polarity classes namely, POS (Positive polarity), NEG (Negative polarity) and NEU (Neutral polarity). In cases of disagreement, the corresponding data was forwarded to a third evaluator, whose judgment determined the final label of the data point. The results obtained are presented in Table 1, which shows the distribution of examples across different classes. It is observed that the majority of the available examples belong to the positive class, accounting for 56 % of the total. The negative class follows with 24 % of the examples, and the neutral class represents 20 % of the samples.

The agreement between the first two evaluators was recorded in 80 % of the cases while it is worth noting that the majority of cases of disagreement were related to the classification of neutral sentiment, accounting for 62 % of the total disagreements. This highlights the additional challenge posed by the inclusion of the neutral class, further complicating an already difficult problem.

3.2. Preprocessing tasks

The sentiment analysis process operates at both sentence and word levels. Therefore, a step is required to tokenize the sentences into words before proceeding further. Additionally, it is good practice to remove special characters and digits, as well as converting the text to lowercase. There are several steps involved in preparing the text for sentiment analysis, and the majority of these steps are implemented in the preprocessing stage. Table 2 presents an overview of the preprocessing tasks applied to the textual data in the context of Lexicon based approach.

In the case of ML algorithms, as part of the preprocessing phase, we employ the use of a pre-trained fastText model to embed the documents from the input corpus into a vector space. This technique enables us to capture the semantic representations of the text, facilitating efficient processing and analysis (Grave et al., 2018).

3.3. Feature selection

Feature selection is a crucial step in machine learning. The quality and relevance of the features used can greatly impact the performance of the modeling algorithm. Having too many features can lead to overfitting and increased computational complexity, while having too few features may result in underfitting and a lack of representative information. Hence, in this step, our objective is to carefully select the most informative features so as to ensure optimal model performance and generalization capabilities while also reducing computational complexity. In the proposed approach, we represent the sentences as vectors using the Sentence-BERT (SBERT) technique (Seo et al., 2022).

3.4. Model training

During this stage, the model learns from the labeled training data and adjusts its internal parameters to make accurate predictions or classifications on unseen data. The training process involves optimizing the model's performance by minimizing the difference between its predicted outputs and the actual labels of the training data. In our approach, we employ the technique of Stratified K-fold Cross-Validation for model training. As proposed by Reimers and Gurevych (2019), and Misra et al. (2017), we set K = 10. Cross-validation (CV) is widely recognized and commonly utilized for training and evaluating the performance of classifiers in practical scenarios (Arlot & Celisse, 2010). It entails partitioning the dataset into two distinct subsets: the training set and the validation set. During the training process, the classifier is presented with the training data set, while the validation set is utilized for assessing the model's performance. There are numerous variations of CV, depending on how the dataset is partitioned. One such variation is Stratified Cross-Validation. In Stratified CV, the sampling of instances for each subset is not random. Instead, the samples are chosen in a way that ensures roughly equal proportions of each class in all subsets. This provides a basic safeguard against imbalanced distributions in the training and validation sets. Through the use of cross-validation techniques, such as Stratified CV, we can effectively assess the performance and generalization capabilities of the model. By training the model on one set of data and evaluating it on a separate set, we can gauge its ability to handle new and unseen instances, and mitigate issues of overfitting. In summary, the use of Stratified CV allows us to obtain reliable estimates of the model's performance and ensure its ability to generalize well to unseen data.

Table 1
The distribution of samples across the sentiment classes.

Actual Class	Number of Instances
POS	795
NEG	341
NEU	284

Table 2
Preprocessing steps applied in lexicon-based algorithm.

Preprocessing Step	Description
Basic text transformation tasks	Convert the text to lowercase or uppercase (if it is deemed that no information will be lost with this conversion) or removing accent marks from the text.
Text Tokenization	The text is segmented into sentences or words using simple rules, such as based on the presence of punctuation marks.
Exclusion of Stopwords	The exclusion of stopwords involves removing commonly used words from the text.
Lemmatization	Lemmatization is the process of mapping all different forms of a word to its base form, known as the lemma.

3.5. Model evaluation

One essential step in the evaluation process is, of course, measuring the performance of the developed model. The quality of a model can have multiple interpretations, but the most common interpretation is its performance on unseen data. Success in this phase depends on two factors: the use of appropriate metrics and the correct interpretation of the results. To ensure the accurate evaluation and ranking of text classification algorithms, we utilize several different metrics that complement each other (Beiranvand et al., 2017). This approach aims to cover various aspects of algorithm performance, including efficiency, reliability, and overall of the solution provided. The selected metrics enable us to assess the effectiveness and robustness of the algorithms, ultimately facilitating the determination of their overall performance. Below, we briefly present the chosen metrics that are recorded throughout this study.

In the case of binary classification, this matrix represents the categories of successes and failures in a 2×2 grid as shown below in Table 3. This matrix provides information about whether certain classes tend to be quality confused with other classes and consequently enables the calculation of the performance metrics (Room, 2019), presented in Table 4.

In our experiment, we are dealing with three classification classes, which correspond to the Positive, Negative, and Neutral sentiments. Thus, in the case of 3 classes, the confusion matrix becomes a square 3×3 matrix, where the (i,j) element represents the number of instances that belong to class i but are classified as class j (Grandini et al., 2020; Room, 2019). It follows that the elements on the main diagonal correspond to the correct classifier decisions, while the elements outside the diagonal represent the errors (or confusion) of the classifier. Capitalizing on the above, we proceed to calculate the metrics outlined in the table that follows.

Table 3
Confusion matrix for binary classification.

		Predicted Category	
		Negative	Positive
Actual Category	Negative	True Negative (TN) (True negative cases: The number of true negative predictions)	False Positive (FP) (False positive cases: The number of false positive predictions)
	Positive	False Negative (FN) (False negative cases: The number of false negative predictions)	True Positive (TP) (True positive cases: The number of true positive predictions)

Table 4
Classification evaluation metrics.

Metric	Calculation	Description
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	Accuracy indicates the percentage of correct predictions made by the classifier. It provides a general assessment of the model's performance across all classes. However, accuracy alone may not be sufficient if the dataset is imbalanced or if different classes have varying degrees of importance.
Precision	$\frac{TP}{TP+FP}$	A model with low precision is capable of identifying a substantial number of positive examples, but it also exhibits a high rate of false positives, incorrectly classifying many non-positive instances as positive. Conversely, a model distinguished by high accuracy may not succeed in capturing all positive examples, but the instances it categorizes as positive are highly likely to be true positives.
Recall	$\frac{TP}{TP+FN}$	Recall is a measure that reflects the ability of a classification model to correctly detect positive instances. A model with high recall performs well in detecting positive instances in the data, although it may also incorrectly classify certain negative instances as positive. Conversely, a model with low recall is unable to identify all (or at least a significant portion of) the positive instances.
F1-Score	$\frac{2 * Precision * Recall}{Precision + Recall}$	It is used to balance the results of Accuracy and Recall into a single measurement of algorithm performance. Therefore, it can effectively capture the model's performance even in cases where the data exhibits significant class imbalance. When the F1 score is in the middle range, it indicates that one of the two components (precision or recall) is relatively low, but the metric does not provide a precise breakdown of which one.

4. Implementation

As previously mentioned, we design our models using the Orange Data Mining platform. Orange is an open-source platform that enables the implementation of data mining processes, supporting a variety of methods from the fields of statistics and machine learning (Demšar & Zupan, 2013). Additionally, the platform complements these methods with interactive visualization and data exploration tools for the data being processed.

4.1. Lexicon based approach

To implement this method, we start by presenting the text data in the workflow as it shown in Fig. 2. Next, we choose to create the corpus which consists of student reviews, and additionally define the dependent variable.

The next steps of our workflow, following our paradigm, consist of the text preprocessing tasks that have been presented in Table 2.

After the preprocessing steps the text polarity is being extracted using the VADER lexicon. We judged that we should rely on the VADER lexicon which is fully open-sourced under the MIT License, in the English language, fully validated and widely accepted across multiple distinct domains including the field of education (Navaneetan et al., 2024; Shafana & Safnas, 2022; Wook et al., 2020). Furthermore, the VADER sentiment lexicon has been compared to many other well-established sentiment analysis lexicons such as Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), WordNet and the Hu-Liu, with quite remarkable results. The comparison results revealed that the VADER lexicon performs at a similar level to that of individual human raters. In fact, the classification accuracy metric "Precision" varied between 0.69 and 0.99 across various domain contexts (Hutto & Gilbert, 2014). At the same time, it was essential that the students express themselves in their mother tongue, Greek, in order for their responses to convey meaning and sentiment as accurately as possible (Sharma et al., 2014). Therefore, to achieve maximum accuracy and efficiency, all the comments were initially collected in the Greek language and subsequently translated into English by accredited translators and interpreters of Greek and English. In this way, we ensured the highest degree of accuracy both from the students' feedback in their mother tongue, as well as by utilizing the VADER English dictionary which provided us with the validity needed for the purposes of our research. We then entered the final data into our system.

The compound score resulting, indicates the overall intensity of sentiment in each sentence. Fig. 3 shows a representative sample of comments with the corresponding compound (sentiment) score for each. The compound score of a specific response is determined by summing up the sentiment scores of each individual word within the sentence and falls within the range of -1 to $+1$ (Qi & Shabrina, 2023).

4.2. Random forest based approach

In the second implemented approach (Fig. 4), we extract the sentiment carried by the text by applying a machine learning method. Our goal here is to train a classifier based on the random forest algorithm and subsequently gather and analyze the performance results. In the initial stage of the workflow, the file containing the comments is loaded into the workflow, and the data is propagated to the steps of the workflow that facilitate the specification of the dependent variable. Subsequently, the text corpus is created.

Prior to presenting the data to the machine learning algorithm, a transformation occurs whereby the comments in the corpus are converted into vectors using the Multilingual SBERT (Sentence BERT) approach (Feng et al., 2020). At this stage, in order to examine the significance of each feature generated by the SBERT method, we rank all features using the Gain and Gini metrics (Tangirala, 2020) and keep the top n features. This particular step allows us to exclude features that do not "explain" the dependent variable and therefore do not further enhance the quality of node separation in the decision trees generated by the method. In the experiments conducted, we did not observe any improvement in the results of the method when selecting more than ninety features. We choose to train and evaluate the random forest model using the Stratified K-Fold Cross Validation technique. This choice is based on our initial observation regarding the uneven

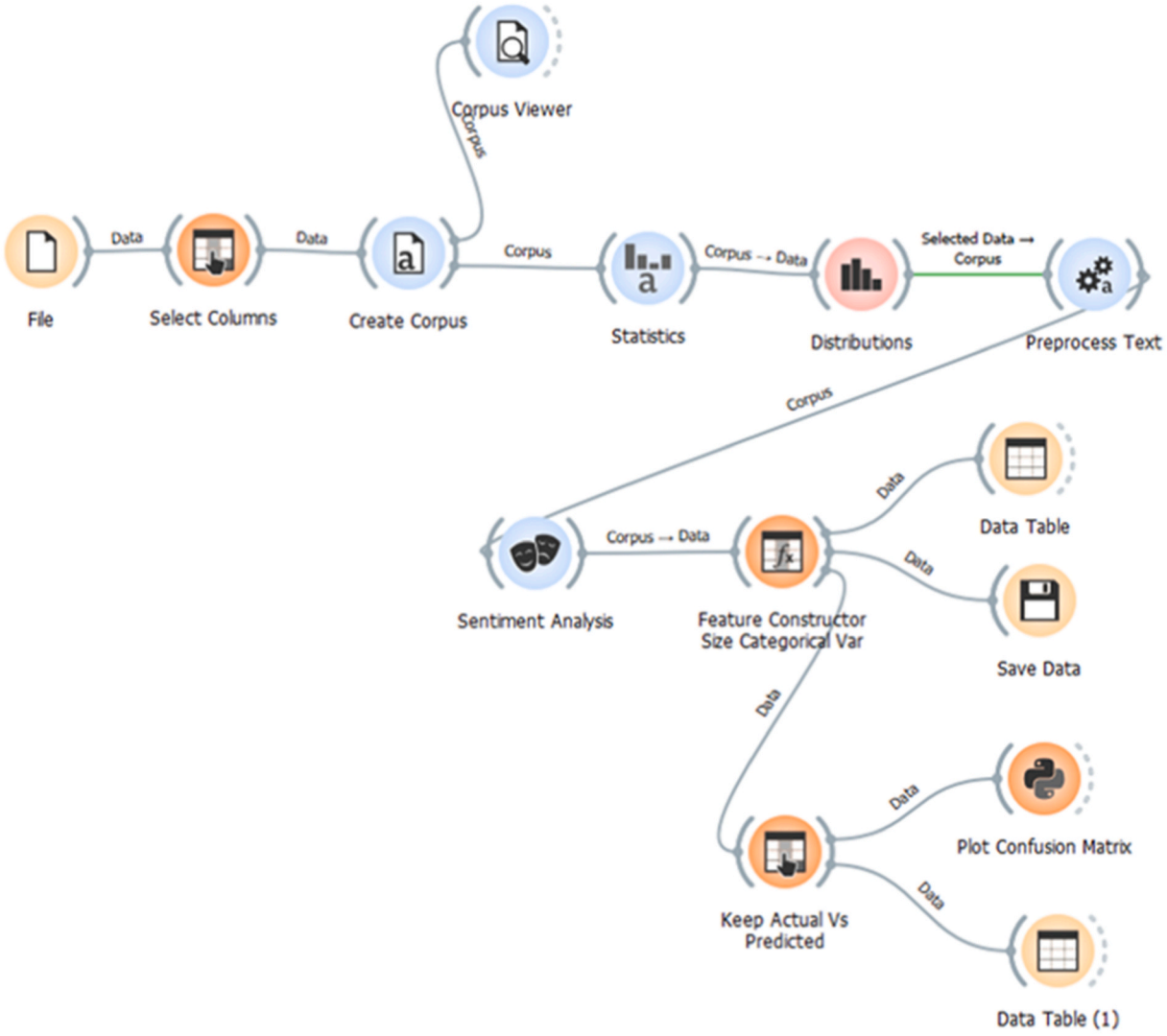

Fig. 2. Lexicon based approach.

Students Comment True	pos	neg	neu	compound
I believe he presents the lesson ...	0.175	0.066	0.758	0.3182
I think that students pay attenti...	0.108	0.072	0.821	0.1779
Your active listening capability ...	0.387	0.075	0.538	0.6997
His persuasion skills are satisfa...	0.148	0.076	0.776	0.2115
he has told us very politely that...	0	0.076	0.924	-0.1027
The professor uses everyday to...	0.176	0.088	0.736	0.3678
His calm way of teaching enco...	0.333	0.09	0.577	0.5859

Fig. 3. Sentiment score of student feedback comments.

distribution of sentiment classes in our data. Furthermore, we set K = 10, an empirically proven value that often yields good results (Anguita et al., 2012).

4.3. Artificial neural network based approach

Similar, to the first machine learning method, the ANN-based method, follows the same preprocessing steps. After creating the corpus and transforming the comments into vectors using the S-BERT technique, the features are ranked based on the Information Gain metric using the "Rank" component. The top sixty features are selected for further analysis. This selection was determined through experimentation, aiming to find the simplest model that maintains a good balance between performance and generalization, effectively reducing the risk of overfitting.

At the final steps of the workflow, the model is trained and evaluated

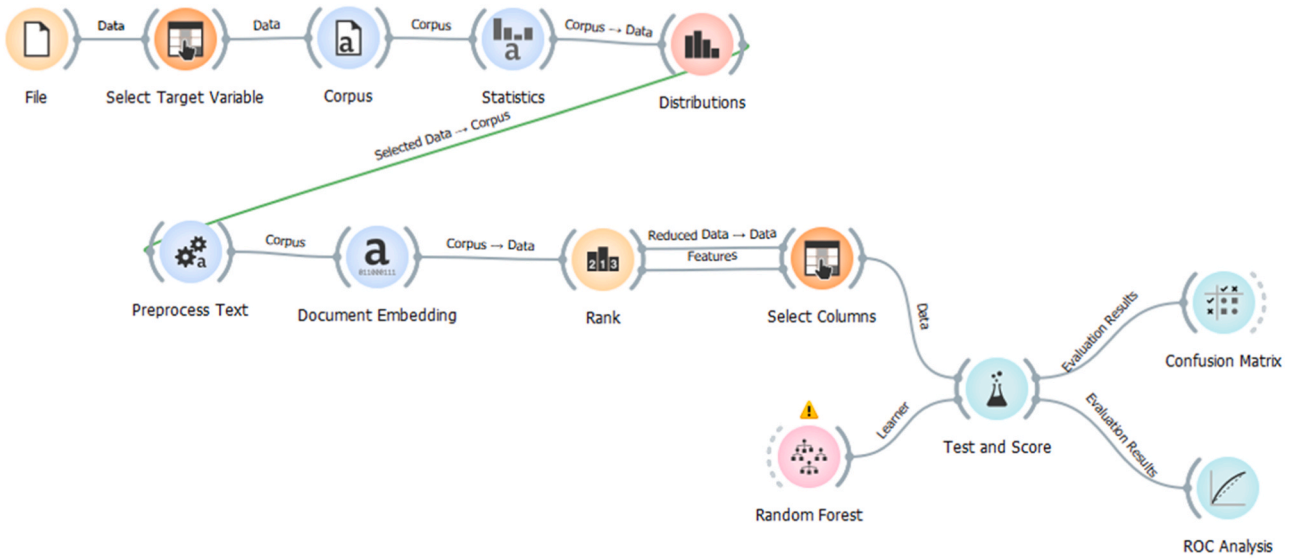

Fig. 4. Random forest based approach.

using the Stratified CV technique. Through an iterative experimentation process, we explored different combinations of hyperparameters to identify the best performing configuration for our ANN model. The resulting hyperparameters are presented in Table 5.

5. Results and discussion

In this section, we comparatively examine the outcomes of the previously implemented methods with the aim of drawing final conclusions, as well as identifying limitations and prerequisites.

As a general observation (Table 6), we note the superiority of machine learning methods over the lexicon-based approach. This observation holds true both in terms of overall accuracy and individual evaluation metrics.

However, it should be emphasized that the performance of the lexicon-based technique is directly dependent on the style of the analyzed text and the specific lexicon used.

Staying within the lexicon-based technique and delving into our analysis of the degree of granularity of each class, we observe (Table 7) that the model's performance for the POS class is relatively satisfactory, with a Recall of 71.9 % and correctly classifying 572 out of 795 comments, the Precision for the same class is 76 %, as out of the 752 responses classified as POS, 572 actually belong to that class. However, the model's performance is not at the same level when it comes to classifying negative (NEG) and especially neutral (NEU) responses, where we observe Recall of 57 % and 38 % respectively.

These values indicate the model's difficulty in recognizing patterns belonging especially to the neutral sentiment class. The same pattern can also be observed in the precision metric. Specifically, as depicted in Table 8, the precision is considerably low in the NEU class. Out of the 291 responses predicted to belong to the NEU class, only 108 were correctly classified, resulting in a precision of 37.1 %.

Focusing on the two machine learning algorithms, we observe that they demonstrate satisfactory performance, with notable differences in capturing the neutral sentiment. Specifically, as shown from Table 10, the ANN consistently performs well in identifying responses from all classes, unlike the random forest, which struggled to capture correlations in the data for neutral comments. More specifically, as observed from Table 9, the random forest algorithm achieves an overall accuracy

Table 5
ANN hyperparameters.

Hyper-Parameter	Value
Neurons in hidden layers	60
Activation Function	Relu
Solver	Adam
Regularization, α	0.07
Maximal number of iterations	2500

Table 6
Evaluation results over all classes.

SA Method	CA	F1	Precision	Recall
Lexicon	0.616	0.552	0.549	0.556
Random Forest	0.734	0.667	0.718	0.662
Artificial Neural Network	0.782	0.759	0.763	0.759

Table 7
Evaluation results by class.

SA Method	Precision			Recall		
	POS	NEG	NEU	POS	NEG	NEU
Lexicon	0.760	0.517	0.371	0.719	0.571	0.380
Random Forest	0.747	0.701	0.706	0.933	0.583	0.350
Artificial Neural Network	0.829	0.701	0.761	0.819	0.819	0.640

Table 8
Lexicon confusion matrix.

	Actual Data				
Classifier results		POS	NEG	NEU	sum
	POS	572	89	91	752
	NEG	97	195	85	377
	NEU	126	57	108	291
	sum	795	341	284	1420

Table 9
RF confusion matrix.

	Actual Data				
Classifier results		POS	NEG	NEU	sum
	POS	742	116	135	993
	NEG	37	199	48	284
	NEU	16	26	101	143
	sum	795	341	284	1420

of nearly 74 % (1042 out of 1420 instances) in correctly predicting the sentiment class. This performance is considered satisfactory given the difficulty of the problem (Roebuck, 2012). .

Analyzing the Recall and Precision metrics for each sentiment class, we observe excellent results for the POS and NEG classes, while it is evident that there is difficulty in identifying instances belonging to the neutral class. In the case of the ANN algorithm, we can conclude that the resulting model shows overall satisfactory results. It achieves a classification accuracy of 78 %, correctly assigning 1112 out of 1420 comments to their respective classes. The ANN-based classifier performs exceptionally well in identifying positive (POS) and negative (NEG) comments, with a Recall of 82 % in both classes.

In other words, the ANN correctly predicts 649 out of 795 positive cases and 281 out of 341 cases belonging to the NEG class. These results are combined with good precision for both classes. Specifically, the model achieves a Precision of 83 % for the positive class, 70 % for the negative class and 76 % for the neutral class. Even in the recognition of neutral messages the ANN observes much better results than the previous classifiers. Specifically, while the Lexicon-based and the RF-based classifier had 38 % and 35 % Recall respectively, the ANN classifier achieved 64 % Recall.

Within the scope of this study, the RF and ANN machine learning algorithms clearly outperform the lexicon-based methods, as Srivastava et al. (2022), highlights in their study, with the ANN particularly excelling in performance across all sentiment classes. Notably, as an extension of the Wankhade et al. (2022) study, the limitation in the case of the ANN is the availability of a large volume of training data and the risk of overfitting the model to the training data, leading to poor generalization ability. Furthermore, as a continuation of the Adeniyi et al. (2022) study, we observed that the integration of the ANN method combined with sentence embedding, outperformed the other approaches. Table 11 summarizes the pros and cons of each method.

These findings are fully in line with the Adeniyi et al. (2022) study, which thoroughly examined and compared machine learning algorithms, however, limited to 2 classes: positive and negative. Even though recent review papers (Leelawat et al., 2022; Mourtzis et al., 2021; Sen et al., 2017) have also examined various sentiment analysis approaches in other domains, none have ventured into the domain of education as from our perspective. Although our study is focused on how the teaching process as well as the teachers are evaluated (identifying and attributing the intensity of sentiment that emerges from a comment regarding their teacher) from the students' point of view, it could be combined, embedding the ANN algorithm which addresses the textual data challenges, with a powerful study conducted by Del Gobbo et al. (2023). In particular, Del Gobbo et al. (2023), in a rather interesting paper, proposed a quite promising framework of the automatic grading of students' answers for exams. Their solution was conceived to offload teachers and instructors from the potentially huge task of grading exam answers to open-ended questions. In summary, our contribution extends prior research by directly applying sentiment analysis to student textual data, comparing different approaches in a challenging 3 class context, positive, negative and neutral, opening up new avenues for further research and discussion, offering opportunities to refine and improve the learning process, the assessment of teacher competencies etc. based on student opinions.

Table 10
ANN confusion matrix.

	Actual Data				
Classifier results		POS	NEG	NEU	sum
	POS	649	54	79	782
	NEG	97	281	23	401
	NEU	49	8	182	239
	sum	795	341	284	1420

Table 11
Pros and Cons of each method examined in our study.

Method	Pros	Cons
Lexicon-based Approach	1. Initial simplicity in implementation. 2. Potential usefulness under specific conditions.	1. Limited in handling the introduction of the neutral sentiment class. 2. Struggles in recognizing patterns within this class, particularly evident in the lexicon-based approach.
Random Forest (RF)	1. Demonstrates the ability to model both linear and complex non-linear relationships. 2. Performance improvement potential with a larger volume of data.	1. Faces challenges in recognizing the patterns of the neutral sentiment class. 2. Interpretability might be a concern.
Artificial Neural Network (ANN)	1. Superior overall performance, especially with a large volume of data. 2. Ability to model complex relationships.	1. Interpretability is compromised. 2. The need for a substantial dataset for optimal performance.

6. Limitations

This study was limited to examining and comparing two different machine learning algorithms, Artificial Neural Network (ANN) and Random Forest (RF) to a specific lexicon (VADER) for sentiment analysis on student textual data. This limited scope of algorithms and lexicon examined, may not be representative or cover the diversity of machine learning algorithms or other lexicon-based methods that may be utilized. However, the selection of the approaches we compared, from which the results were derived, is not exhaustive but covered a substantial range of significant algorithms. In addition, the dataset used, was generated from an educational domain, therefore, the findings may not be entirely generalized to other domains. During the data collection process, the study was also limited to gathering textual data from the student responses of our university in 2 courses and 2 professors. Although obtaining an extensive collection of student feedback encompassing various courses with distinct characteristics, such as different semesters, cognitive areas, and varying degrees of course difficulty, may demand extra time and effort, we are confident that it has the potential to unveil significant insights and conclusions.

7. Conclusion

In our research, we used student comments obtained from feedback at the end of an academic semester and implemented three methods of sentiment analysis to extract the underlying polarity of each comment. Initially, we examined a Lexicon-based method, followed by corresponding implementations based on machine learning algorithms, random forest and artificial neural network. It was observed that the Lexicon-based sentiment analysis technique faced challenges in capturing the subtle nuances of the language. It is worth noting that this method did not necessitate complex preprocessing stages, as required by machine learning methods, because we relied on the existence of updated lexicons that allow for accurate inference. More specifically, the Lexicon-based approach struggle to correctly extract meaning in cases where word order and contextual meaning need to be taken into account. In our case, the text rarely included explicit negative words or descriptors. Instead, the negative tone was conveyed through the entirety of the sentence, making it challenging for the Lexicon-based approach to accurately detect the polarity. In contrast, the machine learning approaches were more effective in their analysis due to their ability to capture the whole meaning of the comment. In short, the linguistic characteristics of the text favored the machine learning approaches over the Lexicon-based method, while the results documented, emphasize the increased difficulty introduced by the inclusion of the

neutral sentiment class to the problem. All classifiers struggled to recognize patterns within this class, particularly with the Lexicon-based approach.

Another key finding of the research is that the machine learning algorithms examined performed better in the specific domain of student feedback responses. Both artificial neural networks and random forests have the ability to model linear as well as complex nonlinear relationships. However, the ANN algorithm demonstrated better overall performance, which can be even further improved compared to RF, utilizing large volumes of data. We should also highlight the pivotal role of the SBERT in achieving more accurate results. This embedding technique allowed the machine learning algorithms to capture the semantic meaning and contextual relationships between the comments more precisely while also speeding up the training process. Therefore, the integration of the ANN method combined with sentence embedding, outperformed the Lexicon-based approach as well as the RF method.

We plan to conduct extensive research on various methods of measuring and leveraging student feedback using and comparing different machine learning algorithms and hybrid approaches adapted to the field of assessment in education. We strongly believe that through our research we can contribute to how institutions face the challenge of harnessing this untapped textual data. Ultimately, there is no intention to undermine or challenge any other approach, rather, we aim to highlight the diversity of approaches and to encourage the exchange of views for the purpose of exploring the potential of this specific context.

Originality and submission status

The authors confirm that this work is original and has not been published elsewhere.

CRediT authorship contribution statement

Panos Fitsilis: Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Charalampos Dervenis:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology. **Giannis Kanakis:** Writing – original draft, Visualization, Validation, Software, Investigation, Data curation.

Conflict of interest

The authors declare no conflict of interest.

References

Adeniyi, J. K., Adeniyi, A. E., Oguns, Y. J., Egbedokun, G. O., Ajagbe, K. D., Obuzor, P. C., & Ajagbe, S. A. (2022). Comparison of the performance of machine learning techniques in the prediction of employee. *ParadigmPlus, 3*(3), 1–15.

Altrabsheh, Nabeela, Gaber, Mohamed Medhat, & Cocea, Mihaela (2013). SAE: sentiment analysis for education. *International Conference on Intelligent Decision Technologies* (Vol. 255,, 2013.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The'K'in K-fold cross validation (April) *In ESANN*, 441–446.

Arlot, Sylvain, & Celisse, Alain (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79. https://doi.org/10.1214/09-SS054

Beiranvand, V., Hare, W., & Lucet, Y. (2017). Best practices for comparing optimization algorithms. *Optimization and Engineering, 18*(4), 815–848.

Chakraverty, S., & Saraswat, M. (2017). Review based emotion profiles for cross domain recommendation. *Multimedia Tools and Applications, 76*, 25827–25850.

Data growth worldwide 2010–2025 | Statista. (2023, November 16). Statista. (https://www.statista.com/statistics/871513/worldwide-data-created) (accessed January, 2024).

Del Gobbo, E., Guarino, A., Cafarelli, B., & Grilli, L. (2023). GradeAid: A framework for automatic short answers grading in educational contexts—design, implementation and evaluation. *Knowledge and Information Systems*, 1–40.

Demšar, J., & Zupan, B. (2013). Orange: Data mining fruitful and funa historical perspective. *Informatica, 37*(1).

Dervenis, C., Fitsilis, P., Iatrellis, O., & Koustelios, A. (2024). Assessing teacher competencies in higher education: A sentiment analysis of student feedback. *International Journal of Information and Education Technology* (vol. 14,(4), 533–541, 2024.

Dervenis, C., Fitsilis, P., & Iatrellis, O. (2022). A review of research on teacher competencies in higher education. *Quality Assurance in Education, 30*(2), 199–220.

Dervenis, C., Kyriatzis, V., Stoufis, S., & Fitsilis, P. (2022). Predicting students' performance using machine learning algorithms (September) *Proceedings of the 6th International Conference on Algorithms, Computing and Systems*, 1–7.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N.& Wang, W. (2020). Language-agnostic BERT sentence embedding. arXiv preprint arXiv:2007.01852.

Gamallo, P.& Garcia, M. (2014, August). Citius: A naive-bayes strategy for sentiment analysis on english tweets. In Proceedings of the 8th international Workshop on Semantic Evaluation (SemEval 2014) (pp. 171–175).

Ganganwar, V., & Rajalakshmi, R. (2023). Enhanced hindi aspect-based sentiment analysis using class balancing approach. *International Journal of Information Technology, 15*(7), 3527–3532.

Garg, K. (2020). Sentiment analysis of Indian PM's "Mann Ki Baat. *International Journal of Information Technology, 12*(1), 37–48.

Gautam, G., & Yadav, D. (2014). Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis (August). *2014 Seventh international conference on contemporary computing (IC3)* (pp. 437–442). IEEE, (August).

Grandini, M., Bagli, E.& Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.

Hardeniya, T., & Borikar, D. A. (2016). Dictionary based approach to sentiment analysis-a review. *International Journal of Advanced Engineering, Management and Science, 2*(5), Article 239438.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews (August) *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge discovery and data Mining*, 168–177.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text (May) *Proceedings of the International AAAI Conference on web and Social Media* (Vol. 8,(No. 1), 216–225.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with Many Relevant Features (April). *European conference on machine learning* (pp. 137–142). Berlin, Heidelberg: Springer Berlin Heidelberg, (April).

Kausar, S., Huahu, X., Ahmad, W., & Shabir, M. Y. (2019). A sentiment polarity categorization technique for online product reviews. *IEEE Access, 8*, 3594–3605.

Leelawat, N., Jariyapongpaiboon, S., Promjun, A., Boonyarak, S., Saengtabtim, K., Laosunthara, A., … Tang, J. (2022). Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning. *Heliyon, 8*(10).

Liu, B (2012). Sentiment analysis and opinion mining. *Synthesis lectures on Human Language Technologies, 5*(1), 1–167.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical biophysics, 5*(4), 115–133.

Mejova, Y. (2009). *Sentiment Analysis: An Overview*. University of Iowa, Computer Science Department.

Misra, Amita, Brian Ecker, and Marilyn A. Walker. "Measuring the similarity of sentential arguments in dialog." arXiv preprint arXiv:1709.01887 (2017).

Misuraca, M., Scepi, G., & Spano, M. (2021). Using Opinion Mining as an educational analytic: An integrated strategy for the analysis of students' feedback. *Studies in Educational Evaluation, 68*, Article 100979.

Mourtzis, D., Angelopoulos, J., Siatras, V., & Panopoulos, N. (2021). A Methodology for the Assessment of Operator 4.0 Skills based on Sentiment Analysis and Augmented Reality. *Procedia CIRP, 104*, 1668–1673.

Navaneetan, M., Tharagesh, G., & Sai Sabitha, A. (2024). Unveiling Sentiments: Analyzing Learner's Experience Using VADER and RoBERTa Models. *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. IEEE,.

Pang, B., Lee, L.& Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: The students' perspective. *Assessment & Evaluation in Higher Education, 33*(2), 143–154.

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach. *Social Network Analysis and Mining, 13*(1), 31.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Roebuck, K. (2012). *Sentiment analysis: High-impact strategies-What you need to know: Definitions, adoptions, impact, benefits, maturity, vendors*. Emereo Publishing,.

Room, C. (2019). Confusion Matrix. *Mach Learn, 6*, 27.

Salmony, M. Y., Faridi, A. R., & Masood, F. (2023). Leveraging attention layer in improving deep learning models performance for sentiment analysis. *International Journal of Information Technology*, 1–10.

Saraswat, M., Chakraverty, S., & Kala, A. (2020). Analyzing emotion based movie recommender system using fuzzy emotion features. *International Journal of Information Technology, 12*, 467–472.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing surveys (CSUR), 34*(1), 1–47.

Sen, T., Ali, M. R., Hoque, M. E., Epstein, R., & Duberstein, P. (2017). Modeling doctor-patient communication with affective text analysis (October). *2017 seventh international conference on affective computing and intelligent interaction (ACII)* (pp. 170–177). IEEE, (October).

Seo, J., Lee, S., Liu, L., & Choi, W. (2022). TA-SBERT: Token attention sentence-BERT for improving sentence representation. *IEEE Access, 10*, 39119–39128.

Shafana, A. R. F., & Safnas, S. M. (2022). Does technology assist to continue learning during pandemic? A sentiment analysis and topic modeling on online learning in south asian region. *Social Network Analysis and Mining, 12*(1), 65.

Sharma, R., Nigam, S. & Jain, R. (2014). Opinion mining in Hindi language: a survey. arXiv preprint arXiv:1404.4935.

Srivastava, Roopam, Bharti, P. K., & Verma, Parul (2022). Comparative analysis of lexicon and machine learning approach for sentiment analysis. *International Journal of Advanced Computer Science and Applications, 13*(3).

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications, 11*(2), 612–619.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731–5780.

Wei, W. (2023). Understanding and supporting the use of feedback from mobile applications in the learning of vocabulary among young adolescent learners. *Studies in Educational Evaluation, 78*, Article 101264.

Wen, M., Yang, D., & Rose, C. (2014). Sentiment analysis in MOOC discussion forums: What does it tell us? (July) *Educational Data Mining*, 2014.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics, 35*(3), 399–433.

Wook, M., Vasanthan, S., Ramli, S., Razali, N. A. M., Hasbullah, N. A., & Zainudin, N. M. (2020). Exploring students' feedback in online assessment system using opinion mining technique. *International Journal of Information and Education Technology, 10*(9), 664–668.

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences, 181*(6), 1138–1152.