UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Student evaluation of teaching: sentiment analysis of student's comments using transformers models vs. artificial intelligence chats like ChatGPT.

Ángel Alejandro Arcos García
José Luis Valencia Vallejo
Inés Micaela Vega Bolaños

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 06 de diciembre de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN DE TRABAJO DE FIN DE CARRERA

Student evaluation of teaching: sentiment analysis of student's comments using transformers models vs. artificial intelligence chats like ChatGPT.

Ángel Alejandro Arcos García

José Luis Valencia Vallejo

Inés Micaela Vega Bolaños

Nombre del profesor, Título académico   Danny Orlando Navarrete Chávez, MS.

Quito, 06 de diciembre de 2023

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos:	Ángel Alejandro Arcos García
Código:	00212578
Cédula de identidad:	1722488739
Nombres y apellidos:	José Luis Valencia Vallejo
Código:	00212577
Cédula de identidad:	1719362327
Nombres y apellidos:	Inés Micaela Vega Bolaños
Código:	00213706
Cédula de identidad:	1727370585
Lugar y fecha:	Quito, 06 de diciembre de 2023

ACLARACIÓN PARA PUBLICACIÓN

UNPUBLISHED DOCUMENT

RESUMEN

La SET (Student Evaluation of Teaching) es una herramienta fundamental para el mejoramiento continuo de la calidad educativa, permitiendo a los estudiantes expresar sus pensamientos y comentarios sobre sus instructores. Sin embargo, el proceso de evaluación y análisis de los comentarios de los estudiantes es generalmente manual y consume bastante tiempo al departamento encargado del proceso en las universidades. Para abordar este problema, el estudio se enfoca en la aplicación de análisis de sentimiento mediante modelos de machine learning, específicamente modelos transformers, y chats de inteligencia artificial. Tanto los modelos, como los chats de inteligencia artificial permiten realizar tareas de procesamiento natural del lenguaje (NLP), que facilitan la clasificación de los comentarios en distintas categorías. El estudio muestra una comparación entre los resultados obtenidos de tres modelos transformers, a los cuales se les realizó fine tuning con los datos recibidos, versus los resultados de tres chats de inteligencia artificial. Dichos modelos fueron evaluados y comparados bajo distintas métricas, y se obtuvo que los modelos DistilBERT (Modelo transformer) y Google Bard (Chat de inteligencia artificial) tuvieron un mejor desempeño en comparación con el resto de los modelos. En resumen, este estudio detalla una investigación y desarrollo exhaustivo acerca de una posible mejora del proceso de evaluación de la enseñanza en una institución de educación superior, a través del análisis de sentimiento, utilizando métodos avanzados de inteligencia artificial y aprendizaje automático.

Palabras clave: SET, NLP, análisis de sentimiento, modelos transformers, chats de inteligencia artificial, machine learning, DistilBERT, Google Bard, fine tuning

ABSTRACT

SET (Student Evaluation of Teaching) is a fundamental tool for the continuous improvement of educational quality, allowing students to express their thoughts and comments about their instructors. However, the process of evaluating and analyzing student feedback is usually manual and consumes considerable time for the department in charge of the process in universities. To address this problem, the study focuses on the application of sentiment analysis using machine learning models, specifically transformer models, and artificial intelligence chats. These models, as well as the artificial intelligence chats, allow performing natural language processing (NLP) tasks, which facilitate the classification of comments into different categories. The study shows a comparison between the results obtained from three transformer models, which were performed fine tuning with the received data, versus the results of three artificial intelligence chats. These models were evaluated and compared under different metrics, and it was obtained that the DistilBERT (transformer model) and Google Bard (artificial intelligence chat) had a better performance compared to the rest of the models. In summary, this study details a comprehensive research and development about a possible improvement of the teaching evaluation process in a higher education institution, through sentiment analysis, using advanced artificial intelligence and machine learning methods.

Key words: SET, NLP, sentiment analysis, transformers models, artificial intelligence chats, machine learning, DistilBERT, Google Bard, fine tuning

CONTENTS TABLE

TABLES INDEX

FIGURES INDEX

INTRODUCTION

Student evaluation of teaching (SET) in higher education institutions has grown in importance as an assessment tool; nonetheless, it should only be used sparingly because relying too much on it to determine a student's success might exacerbate academic stress. (Cook et al., 2021). The evaluation of faculty can be carried out in a quantitative or qualitative instruments, for this particular case it has been decided to do it in a qualitative way because it allows a better understanding of the feelings and senses of the people and therefore improves the results (Gupta et al., 2020). The problem with teaching evaluation methods that are qualitative such as open-ended question is that they are done manually, i.e., one or several people oversee reading each student's comment and classifying it as positive, neutral, or negative (Shaik et al., 2023). Once the comments have been classified, they are collected and analyzed to provide feedback to teachers to improve their performance in class. It is worth mentioning that the feedback not only focuses on the teachers' performance, but also reveals problems in the pedagogical structure of the class and in the learning practices (Shaik et al., 2023). In addition, the results obtained from the analysis of comments are used for decision making by the authorities of educational organizations (Sproule, 2000).

In educational settings from primary school to university levels there has been a rising adoption of tools and applications powered by artificial intelligence (AI) technologies utilized by both educators and learners (Chen et al., 2020). One of these applications is known as sentiment analysis, and it could help to minimize the time spent on this activity and generate better insights for the organizations. Sentiment analysis or opinion mining is the field of study that analyzes written texts in which people express their opinions, feelings, emotions and attitudes toward entities and their attributes (Mohd et al., 2023). In the educational context, students express their sentiments toward the instructors and various attributes such as their teaching methods, interaction with students, experience and knowledge sharing, behavior, and attitudes. Sentiment analysis has different techniques, and they could be categorized as follows: lexicon-based approach, machine learning approach, hybrid approach and others like transformer learning or aspect-based approach (Pooja & Bhalla, 2022). Each of these categories have more subdivisions but, in this study, the main focus is using transformers networks (machine learning approach) because these models provide a substantial solution to the long-standing problems faced in sequential manipulations, opening avenues at an impressive pace in the NLP research space (Acheampong et al., 2021).

This study aims to determine the best model for sentiment analysis in teaching performance evaluation comments, comparing pre-trained models versus artificial intelligence chats.

DEVELOPMENT OF THE TOPIC

1. Literature review

1.1. Student evaluation of teaching (SET)

The most popular method of determining whether a faculty member in a higher education institution is effective in his or her job at this time in higher education institutions is student evaluations of teaching (Dodeen, 2013). Primarily, this is due to the focus on continuous improvement of educational quality, through feedback from students to instructors (Clayson, 2011). Generally, SET is used after the end of the class, this evaluation tool allows students to express their thoughts, emotions and comments to the instructor, whether it is a positive or negative comment. (Vargas, 2023). Although SET may offer positive feedback for a job well done, most experienced instructors have also encountered evaluations that they believed are unreliable, unfair, or even maliciously done (Rupp, 2023). In addition, administrative, instructional, and research critiques of the use of SETs in higher education as faculty evaluations are on the rise (Santisteban & Egues, 2022). This process then allows educational institutions to identify areas of strength and weakness in teaching, enabling them to make decisions that help improve the quality of education. To make effective use of the time of students and professors, the assessment process must ensure that the data acquired is solid, accurate, and insightful (Gupta et al., 2020). Therefore, it is crucial to analyze and interpret these findings effectively and appropriately to implement beneficial modifications in the academic plan.

1.2. Machine learning approach

Sentiment analysis has different techniques, which could be mainly classified as follows: lexicon-based approach, machine learning approach, hybrid approach and others such as transformer learning or aspect-based approach (Pooja & Bhalla, 2022). The idea and creation of computer systems to carry out operations that ordinarily would need human cognition, including as perception, language understanding, reasoning, learning, planning, and problem solving, is referred to as "artificial intelligence." (Nelson et al., 2020). Different applications of artificial intelligence such as machine learning, natural language processing, and deep learning allow the automation and thus optimization of a wide variety of human tasks (Pooja & Bhalla, 2022). One of the core areas of artificial intelligence (AI) is machine learning. Artificial Intelligence (AI) is essentially the development of systems that are capable of tasks like thinking, problem solving, and decision making that would typically need human intelligence (Haakman et al., 2021). Machine learning is a branch of artificial intelligence that focuses on creating models and algorithms that let computers recognize patterns in data and get better over time at certain tasks. In other words, machine learning is a fundamental component of artificial intelligence (AI) as it facilitates autonomous learning and adaptation, allowing for intelligent decision-making and process automation (Sarker, 2021). These range from finding patterns or correlations in data to clustering and classification models, which will be used to label student comments (Dake & Gyimah, 2023). In this case, since the objective of the teacher evaluation is to classify student comments into categories, some machine learning models such as multinomial naive bayes, K-nearest neighbor, neural networks, support vector machines, etc. could be implemented with the parameters configured based on their statistics (Pooja & Bhalla, 2022). These models are based on sentiment analysis that uses natural language processing and machine learning to systematically extract, quantify, identify, and analyze text information generated by people, both objectively and subjectively (Laifa et al., 2023). Generally, several models are evaluated to find the one with the best performance based on its accuracy in classifying comments (Giang et al., 2020).

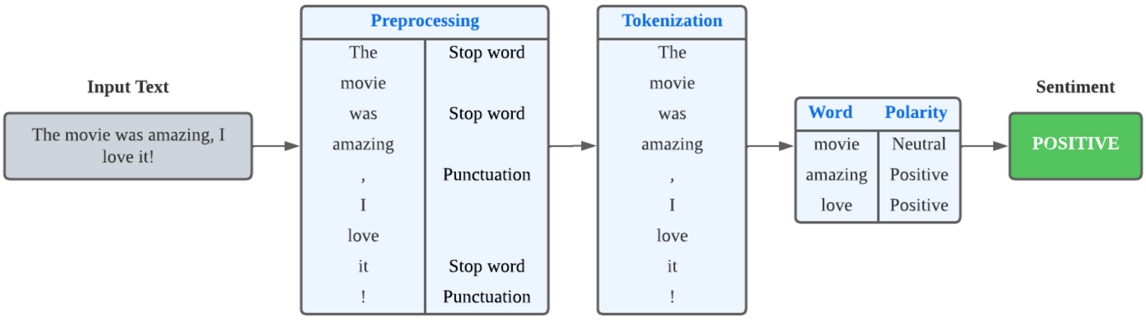
1.3. Transformers models

Transformer models are a particular kind of artificial intelligence technology used in natural language processing. They are based on neural networks, which are a machine learning technique. Mechanism-based transformer models show promise in various natural language processing applications, including sentiment analysis, and transformer models such as BERT, GPT, RoBERTa, help improve NLP tasks (Gheewala et al., 2024). For NLP applications such as text classification, named entity recognition, question answering, language modeling, etc., simple transformer models or lighter variants of transformer models can be created (Mathew & Bindu, 2022). These models consist of encoders and decoders that are stacked together; cross-attention between encoders and decoders is also provided, as is automatic attention to each of the encoded and decoded units (Martínez Hernández et al., 2023). The encoder design of the transformer network manages the symbolic connections between an input categorization and an ongoing relationship. The decoder component of the transformer model, in turn, creates one output sequence after another. Each auto-decoding stage uses the previous input as a complement to the subsequent word (Kotei & Thirunavukarasu, 2023). Mathew & Bindu (2022) mention that the process of building the models is made up of different steps consisting of initializing, training, evaluating, and testing a task-specific model.

One of the advantages of these models is that they have several contributions in different languages, such as English, French, Spanish, and Chinese (Miranda et al., 2023). This means that it is not limited to only one group of people because of the language barrier.

Figure 1

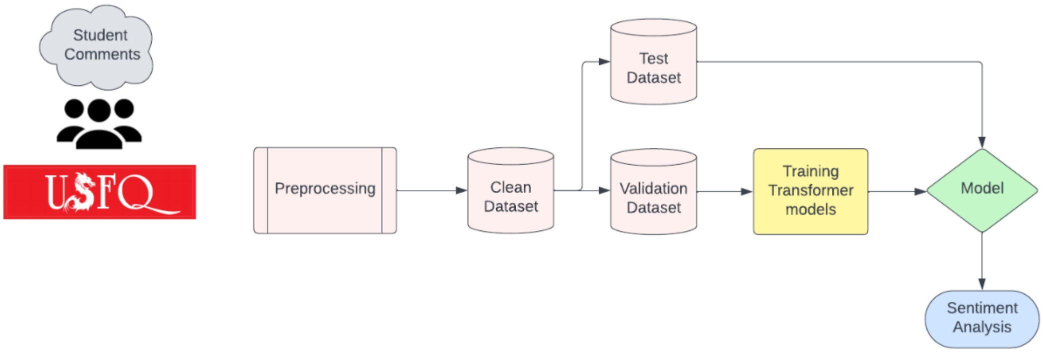
Example of sentiment analysis process based on transformers models



Note. Adapted from "Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. Sustainability" (p. 8), by Ainapure et al., 2023, Sustainability, 15(3), 2573.

Figure 2

General process followed by the models to be used



Note. Adapted from "Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. Sustainability" (p. 13), by Ainapure et al., 2023, Sustainability, 15(3), 2573.

There are different types of Transformers models, but the following models were selected to conduct the sentiment analysis due to their features:

1.3.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a technique that is based on neural networks for pre-training in NLP and the goal of this Google's algorithm is to interpret our search language in a more natural way, using Neuro-Linguistic Programming (Martínez et al., 2023). BERT calculates the embedding of tokens by using bidirectional recurrent neural networks and one of its main advantages is that it considers the context of the term when calculating the embedding, it means that tokens could have different meanings depending on the context (Salazar et al., 2023). In 2018, Devlin et al. defined two steps for this technique: pre-training and fine-tuning. In the pre-training step, it uses two mechanisms: Masked Language Modelling (MLM) and Next Sentence Prediction (NPS) that help to understand language. MLM takes some random sentences as input, masks some of the words in the sentences, and reconstructs the masked words from the context in the output (Acheampong et al., 2021).

1.3.2. RoBERTa

RoBERTa is a Transformers-based language model that was developed with natural language processing (NLP) needs in mind. For particular NLP tasks, the RoBERTa model can automatically extract a variety of features based on context, which aids in capturing the true context of the language (Malik et al., 2023). In order to retrieve key features associated with each word, the RoBERTa model uses an attention mechanism, and it can match these specific connections in the attention heads. (Siddharth & R Aarthi, 2023). RoBERTa is an enhanced variant of BERT. According to several studies the BERT model was insufficiently trained, and in a replication study, RoBERTa achieved better results compared to the BERT model. (Moussa et al., 2022). RoBERTa is allows a critical evaluation of its performance in various languages in comparison with other models (Delobelle, 2020).

The pre-training phase of RoBERTa is where BERT differs most from RoBERTa. RoBERTa has only been pre-trained on one language challenge, not both. RoBERTa scans the complete text input sequence at once and feeds it via a stack of transformer encoders, in contrast to directional models, which read the text input sequentially (Riadh Meghatria et al., 2020). RoBERTa is trained with a substantially bigger amount of textual material and in numerous pre-training steps rather than merely being given the task of filling in holes in phrases. This includes activities like predicting a mask or the following statement, among others. Additionally, RoBERTa modifies significant hyper-parameters and incorporates extra training data.

1.3.3. DistilBERT

One of the most widely used transformer models for classification problems is DistilBert. A faster, lighter, and easier-to-fine-tune variation of BERT is the DistilBERT architecture (Kaminska et al., 2023). By using only half of the training parameters of BERT model, DistillBERT employs a distillation process to filter the training parameters. It runs 60% faster than a standard BERT and keeps 97% of the language capabilities. The key concept is that when a big neural network trains an output, a smaller network can be used to approximate it. (Bakare et al., 2023). This is why DistilBERT is one of the models chosen for the classification of student comments.

1.4. Artificial intelligence chats

To compare the machine learning models (transformers models), three artificial intelligence chats were chosen that are widely used today where many NLP tasks can be performed in which the user interacts directly with the chat.

1.4.1. ChatGPT-4

A pre-trained generative transformer (GPT) is an NLP algorithm that generates human-like text from text input (Radford, 2023). Using linguistic autoregression, GPT models are trained to predict subsequent tokens based on all previously identified tokens (Martínez, 2023). From a technical point of view, ChatGPT is based on OpenAI's pre-trained Generative Pre-trained Transformer (GPT) language models, in particular the GPT-3 and

GPT-4 versions. These models are large-scale neural networks designed for language generation and have been trained on a wide variety of texts from the Internet, and their learning has been reinforced by human feedback (Carsten & Eke, 2023).

1.4.2. Google Bard

Google AI created Google Bard, a large language model (Nguyen, 2023). Bard is able to cite sources, translate texts, read images and search on the Internet in real time to obtain up-to-date information. (Huynh, 2023).

The model receives a lot of training data during this process, and it gains the ability to recognize patterns in the data (Aydin, 2023). The model gets more proficient at producing text, translating between languages, creating other kinds of original material, and responding to queries as it gains knowledge (Moons, 2023).

Google Bard is a versatile tool for creating content because of its ability to produce text in a variety of voices and registers (Rahaman, 2023). Although it is still in development and has limitations and difficulties in its use. Its quality and accuracy may be affected by lack of information or unusual updates (Gan, 2023).

1.4.3. Bing Chat

Microsoft Bing Chat can respond to almost any query or question made by the user, and your answer will be provided from reliable sources. (Pustikayasa, 2023). Bing Chat interprets users' natural language and responds with precise and beneficial information by utilizing machine learning and natural language processing technology (Xuan-Quy, 2023) Users can inquire about anything, from online product searches to weather information. In addition to offering suggestions for goods and services, Bing Chat can assist users in finding details about nearby events and locations (Caramacion, 2023).

1.4.4. Prompt engineering

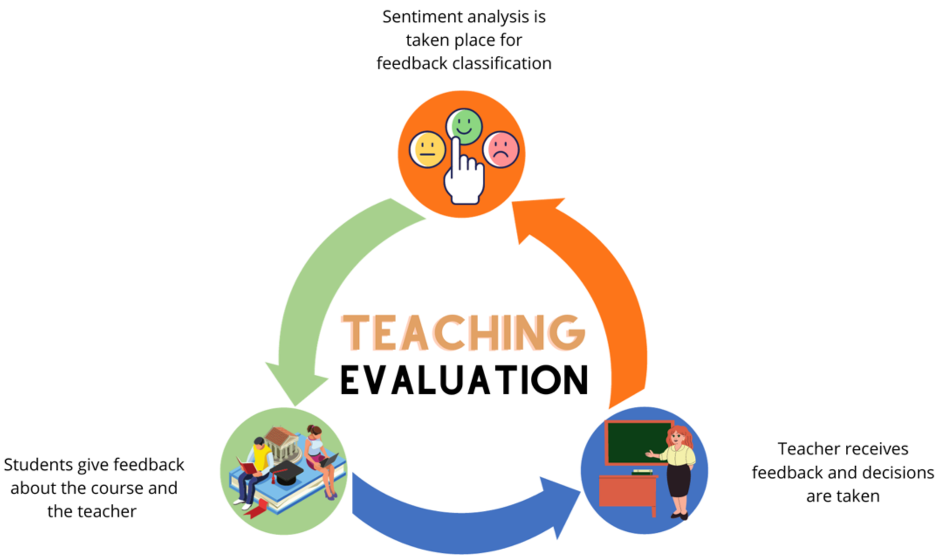
A prompt is considered as a set of instructions that help to custom program an LLM (large language model) by enhancing or refining its capabilities (Liu et al., 2023). There are several applicable techniques depending on the objective pursued. The technique selected to accomplish sentiment analysis was Few-Shot Learning (FSL) because it is characterized by the ability of machine learning models to generalize based on a small number of training examples (Parnami & Lee, 2022).

1.5. Sentiment analysis

Sentiment analysis could be defined as “the computational study of people’s opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes” (Liu, 2015). This analysis can be categorized as a classification tool that comes from Natural Language Processing (NLP), that is the research discipline focused on the automatic processing of human language (Gottipati et al., 2018). Sentiment analysis has wide applicability in various areas such as understanding users' opinions on a product, travelers' feelings, and identifying the polarity of attitudes in tweets (Lazrig & Humphreys, 2022). Figure 1 shows a simplified representation of how sentiment analysis works in relation to the feedback provided by students to faculty members.

Figure 3

Student evaluation of teaching representation


Sentiment analysis is taken place for feedback classification
TEACHING EVALUATION
Students give feedback about the course and the teacher
Teacher receives feedback and decisions are taken

Note. Adapted from "Using sentiment analysis to evaluate qualitative students' responses" (p.4), by Dake, D. & Gyimah, E., 2023, Education and Information Technologies, 28(4), 4629–4647.

1.5.1. Sentiment analysis tasks

Sentiment analysis has different tasks that could be considered as steps that should be followed, there are five main tasks (Wankhade et al., 2022):

1. Sentiment classification: Is a well-known categorization task in this field and it has three subtasks:
   a. Polarity determination: In this step, the sentiment of the unit of analysis is determined, characterized by two key aspects: polarity, indicating positive or negative values, and intensity, which refers to the range of these values. (Tian et al., 2018). It is important to emphasize that neutral values are also usually included.
   b. Cross-domain determination: It is the task in charge of predicting the sentiment of a target domain.

    c. Cross-language determination: This analysis is similar to the previous one, but it aids in understanding text information across different languages.

2. Subjectivity classification: Identifies subjective cues, emotional expressions, and subjective thoughts.
3. Opinion spam detection: It is about recognizing fake reviews that promote or criticize a product for their benefit.
4. Implicit language detection: Implicit language can include some aspects such as sarcasm, irony and humor that can completely change the meaning of a sentence and therefore, its polarity.
5. Aspect extraction: It is a key process where the aspect is extracted, it can be predefined based on the domain being worked on or other approaches can be used such as Frequency-based methods, syntax-based methods, supervised and unsupervised machine learning approaches.

1.5.2. Sentiment analysis levels

Different units of analysis are considered in sentiment analysis, each one with a specific purpose and their respective advantages and disadvantages. According to Ghorbanali & Sohrabi (2023) the most important levels for analyzing sentiments are the following:

- Sentence level: The unit of analysis in this case is the sentence and each one is examined individually where there will only be one polarity.
- Document level: The entire document is considered as a unit and has a negative or positive polarity; however, this level of analysis presents problems because there is not enough detail about all aspects of an entity, and it is not appropriate when there are conflicting emotions that may influence the final decision.

- Aspect level: In this case it is analyzed in a more in-depth way, since the unit of analysis is the entity with their aspects and subsequently, the opinions of each of them are evaluated. For example, from the sentence “iPhone's voice quality is excellent, but its battery is poor”, entity extraction should recognize “iPhone” as the entity, while aspect extraction should identify that “voice” and “battery” are two aspects (Zhang et al., 2018).

1.5.3. Sentiment analysis in education

In Lazrig, & Humpherys (2022), present the results from applying nine machine learning algorithms in different experimental configurations. Some of the algorithms used can be mentioned: Naive Bayes, Decision Tree, Random Forest, Logistic Regression, AdaBoost, etc. Within their results, they obtained 98% accuracy with Naive Bayes when they only had the polarity of positive and negative but excluding neutral, and they also concluded that in an educational context, current algorithms still do not accurately classify neutral sentiments.

Giang et al., (2020) propose to build a system to categorize feedback from Vietnamese university students automatically. They use 3 classifiers: Naive Bayes, Maximum Entropy and Support Vector Machine where they obtain that the Maximum Entropy algorithm is the one that has the higher accuracy (91.36%).

In the paper of Ngoc et al. (2021), BERT were used to predict positive, negative or neutral status of an online course (Coursera) from student reviews. A comparison was made with other algorithms: Decision Tree and SVM and it was found that BERT had a higher accuracy (88.93%).

Having conducted this literature review, the following research questions are formulated:

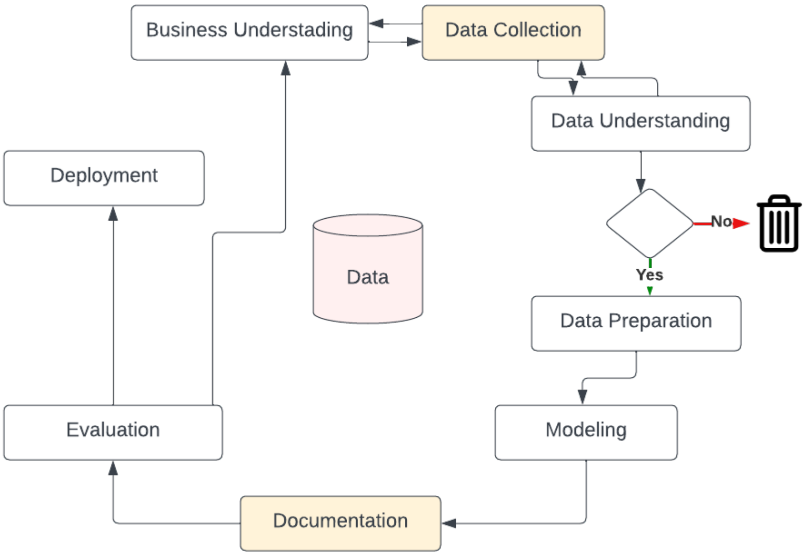
1) Is it possible to use sentiment analysis in an educational context to assist instructors in evaluating students' learning experiences?
2) What is the most suitable model for performing sentiment analysis in an educational context when comparing transformer models with artificial intelligence chat systems?

2. Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was chosen. This methodology is used to improve the success of data mining projects (Moro, 2011). It consists of six iterative phases that go from business understanding to implementation (Huber et al., 2019) The six interconnected steps of the data mining process include business understanding, data understanding, data preparation, modeling, evaluation, and deployment, according to this structured methodology (Christoph Schröer et al., 2021). CRISP-DM offers a strong framework for methodically and effectively handling data analytics projects. An in-depth comprehension of the goals and requirements of the business is the first step, which is followed by the preparation and gathering of data, the creation of descriptive or predictive models, an assessment of the models' performance, and ultimately the integration of the findings into the daily operations of the company. For data teams and data scientists to make well-informed decisions and extract value from data, this methodology encourages an iterative and collaborative approach.

Figure 4

CRISP-DM methodology's representation



Note. Adapted from "DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model" (p. 2), by Huber, S. et al., 2019, Procedia CIRP, 79, 403–408.

The project was carried out at Universidad San Francisco de Quito (USFQ), which is a prestigious Ecuadorian university, but stands out internationally for its academic focus, liberal arts philosophy and constant continuous improvement. The process of collecting, processing and analyzing the student evaluation of teaching at USFQ currently takes about two weeks, is carried out by two people full time and is entirely manual work. In addition, the dataset is from USFQ students that fill in a form with feedback (both quantitatively and qualitatively) about their instructors at the end of each semester. On the other hand, the student input cannot be exploited to their maximum advantage due to the lack of automated text analysis tools. Consequently, the methodology's six steps were implemented, as detailed below.

2.1. Business Understanding

The business understanding section is the first step of this methodology, in this section you must understand the objectives, success metrics and business background (Chapman, 1999). This knowledge must then be transformed into a definition of the problem, in this case data mining, and continue in detail with the rest of the methodology designed to achieve the objectives (Wirth, 2000).

Universidad San Francisco de Quito (USFQ) is a leading institution of higher education in Ecuador and even in Latin America, widely recognized for its academic excellence, constant research, and social action (Delgado, 2012). USFQ has a wide range of programs, exceptional instructors, and state-of-the-art facilities that set it apart as a premier institution in the area. It should be mentioned that USFQ's emphasis on academia and ongoing commitment to improvement are among its primary attributes and advantages.

Continuing with the theme of continuous improvement, USFQ conducts a Student Evaluation of Teaching (SET) every semester, so that students can evaluate the teachers of each subject in different parameters (punctuality, pedagogy, respect, compliance with the syllabus, among others). This evaluation has two components, a quantitative component and an open-ended question component where students can write their comments. Currently, manual work is performed, both for the collection, processing and analysis of the comments given by the students, which is very tedious and time-consuming.

The application of sentiment analysis through transformer models, focused on the classification of data is essential for the university, because it would allow streamlining and automating the current processes because it would save time, human resources and money for any institution of higher education (Giang, 2020).

2.2. Data Understanding

The initial phase of data collecting is followed by actions aimed at familiarizing oneself with the data, identifying issues with its quality, gaining preliminary insights into it, or identifying intriguing subsets (Wirth & Hipp, 2000). For this case, the extraction and collection of student comments is done through the university database.

The database will be used to train the different models; however, this database must go through pre-processing to achieve the expected effectiveness and accuracy (Kalra, 2017). It is also important to note that the database already has classifications (positive, mixed and negative) that were performed manually by the staff of the university's continuous improvement department. The database contained a total of 6,000 comments distributed among the distinct categories. There were 1800 data for each of the positive, mixed, and neutral categories, plus 600 comments classified as "not applicable". Several studies indicate that having a dataset with balanced data helps significantly to have better results, especially in the accuracy of the models (Alshamsi, 2020).

2.3. Data Preparation

This step is essential for the analysis of student comments because some of them contain special characters, emoticons, words and other elements that do not add value to the sentence (Ghosh et al., 2023). Therefore, these aspects must be considered to ensure that the data entered into the models are as clean as possible and do not add noise to the models (Ghosh et al., 2023). On the other hand, machine learning models do not understand human language by itself; they only understand numerical data, so it is necessary to transform comments into numerical data. This is done by tokenization and vectorization, which vary according to the model applied (Krouska et al., 2016).

2.3.1. Text cleaning

First, comments classified as "not applicable" were eliminated because of the lack of meaning and did not contain enough characters to reflect a sentiment or to classify them within the 3 categories mentioned above. On the other hand, comments containing less than 3 characters were eliminated, since these comments added noise to the categories. Thus, in total there were 5317 comments, 1759 from each category, making the datasets balanced.

Several comment-focused preprocessing techniques were applied, starting with lowercase that changes all letters from uppercase to lowercase, and the goal is that the dimensionality of the problem is reduced (Duong & Tram-Anh Nguyen-Thi, 2021). That is, if there are two equal words, one that starts with uppercase and one in lowercase, both have the same dimension. In addition, stop words, non-semantic divisions of natural language that make the text dimensionally larger, and redundant features that are not of interest for classification with the models were removed (Amirita Dewani et al., 2021). On the other hand, it is quite common to use abbreviations and misspellings in the comments, so a manual list of abbreviations and wrong-spelling lexicons was applied based on the most common errors observed in the comments. Furthermore, punctuation marks should be removed as they do not influence the sentiment of the comment and add noise to the models (Duong & Tram-Anh Nguyen-Thi, 2021). Along with punctuation marks, special characters such as "$", "%" and "#", among others, were removed. Finally, numbers were removed from the comments where they were found, as these do not contain any sentiment.

2.3.2. Normalization

Normalization allows to reduce the vocabulary complexity of the proposed models (Kayvan Tirdad et al., 2021), therefore it is a major step for the training of the models. In this case, tokenization was used as a normalization technique, which consists of separating each word individually from the comments and was implemented in the clean datasets (train, validation and test).

2.3.3. Vectorization

An important aspect to consider is that computers do not understand natural human language; words must be transformed into numbers in order to be processed. In other words, it is necessary that a real-number vector representation, also known as word embedding, be applied to the words (Kayvan Tirdad et al., 2021). There are several methods for turning text into vectors such as Bag-of-Words, N-gram, Tf-idf, Word2Vec, GloVe, Doc2Vec, which differ in the way in which they transform words to vectors (Rani et al., 2022). On the other hand, each of the AI models and chats used has its own defined vectorization method, with which the comments were processed after tokenization.

2.4. Modeling

The codes used as a base were extracted from the Hugging Face Hub and adapted according to the data, graphs, and metrics used for the analysis of the models. The codes of the models, preprocessing, and graphs of the confusion matrices of the AI chats can be reviewed on GitHub through the link in Appendix A. In addition, Google Colaboratory was used to make the previously adaptations, and due to limited memory and disk space, the models were run in the Computer Science laboratories of the Universidad San Francisco de Quito. Although the time it took to run all the epochs varied according to the model, each run took about 18 minutes to complete. Therefore, each model required 72 minutes to run. On the other hand, "Hugging Face Hub is a Git-based social code hosting platform focusing on ML development used to host pre-trained ML models. It stores information about the dataset/s and library the models rely on, a widget to run inferences for such model, recommended configuration and spaces that use that model for demo applications" (Ait et al., 2023). To perform the fine tuning of the selected models, the database was divided into train, validation and test sets with 60%, 20% and 20% (Miranda et al., 2023) respectively (3189, 1064, 1064 comments). It is important to define the basic architecture of the transformer models, which is summarized in Table 1.

Table 1

Definitions of architecture parts of transformer models

Architecture Part	Definition
Parameters	Number of learnable variables/values available for the model.
Transformer layers	Number of Transformer blocks. A transformer block transforms a sequence of word representations to a sequence of contextualized words (numbered representations).
Hidden Size	Layers of mathematical functions, located between the input and output, that assign weights (to words) to produce a desired result.
Attention Heads	The size of a Transformer block.

Note. Hugging Face (2018)

It is worth mentioning that BERT and RoBERTa have the same architecture: 12 transforming layers, 768 hidden sizes, 12 attention heads, and 110M parameters (Tang, 2020). The difference is that RoBERTa has different hyperparameters that allow eliminating the next-sentence pre-training objective and training with much higher mini-batches and learning rates (Hugging Face, 2018). On the other hand, DistilBERT differs in that it has six transform layers compared to the other models. It has 40% fewer parameters than Bert-base-uncased and runs 60% faster while preserving more than 95% of the performance of BERT (Hugging Face, 2018). This makes it a small, fast, cheap, and lightweight Transformer model trained by distilling base BERT.

On the other hand, each model was trained with certain hyperparameters mentioned below in Table 2, which were assigned based on similar sentiment analysis work (Sun et al., 2019; Tang et al., 2020; Vásquez et al., 2021), on (Hugging Face, 2018), and on the previous steps of the methodology, data understanding and data preparation. The MAX_LEN was defined since none of the comments exceeded 256 tokens, so a larger number such as 512 was not going to be adequate. On the other hand, for the BATCH_SIZE of the train, validation and test sets, the amount of available data was considered, which was not so large that a high value was not necessary for these hyperparameters. Lastly, the number of EPOCHS and LEARNING_RATE were defined since they were equal values in the investigations mentioned above.

Table 2

Definitions of hyperparameters of models

Hyperparameter	Definition	Value
MAX_LEN	The maximum number of tokens in a sequence to be used for the model.	256 [a]

TRAIN_BATCH_SIZE	The number of data samples used in a training iteration.	4 [a]
VALID_BATCH_SIZE **TEST_BATCH_SIZE**	Like the training batch size, it is the number of data samples used in a validation and test iterations.	2 [a]
EPOCHS	Represents the number of times the model passes through the entire data set during training.	4 [b, c, d]
LEARNING_RATE	It is a critical hyperparameter that controls the magnitude of the adjustments made to the model weights during the training process.	2e-05 [b, c, d]

Note. [a]Hugging Face (2018). [b]Sun (2019). [c]Tang (2020). [d]Vásquez (2021).

Once the training of the models was completed, the artificial intelligence chats were used to compare their results. Both Google Bard and Bing Chat do not offer the option of uploading files for use, so the test set comments were entered and classified one by one. This process was extensive and varied depending on the chat used. In the case of Google Bard, although there was no limit on the number of queries, it stopped working correctly after approximately 50 queries. Therefore, the process had to be stopped for a few minutes to obtain satisfactory results. On the other hand, Bing Chat has a maximum number of 5 queries in a row, so every 5 queries, the instruction to the chat must be repeated. These restrictions influenced the time spent on each AI chat. ChatGPT-4 does offer the option to upload files and use them for different purposes, so the process was significantly faster compared to the other AI chats and even with the trained models. It is worth mentioning that prompt engineering is an aspect that can improve the results obtained, and there are many techniques that can be adapted to different tasks. In this case, the FSL technique was incorporated for each of the artificial intelligence chats that will be used to perform the sentiment analysis: ChatGPT-4, Google Bard and Bing Chat. The prompt consisted of the following: two examples of comments from each category and the sentiment analysis of the rest of the comments (test data set) was requested based on those comments.

2.5. Evaluation

This phase has three tasks to analyze the adaptability of the model to the business: evaluate results, review process and determine next steps (Saltz, 2021). For the first task, it is necessary to define the performance metrics that will be used to evaluate the models. The F1-Score is considered one of the most widely used measures in NLP and machine learning tasks because it combines precision with recall into a single measure (Manias et al., 2023). Additionally, accuracy refers to the extent to which the ML algorithm's predictions of sentiment scores, such as positive, neutral, or negative, align with those of human raters (Lazrig, & Humpherys, 2022). Table 3 summarizes the results obtained for each model evaluated with the previously determined performance metrics (accuracy and F-1 Score).

Table 3

Performance metrics of the models

Models	Accuracy	F-1 Score
BERT	70.7%	0.689
RoBERTa	73.8%	0.741
DistilBERT	**76.5%**	**0.765**
ChatGPT-4	33.6%	0.200
Google Bard	**78.3%**	**0.767**
Bing Chat	74.6%	0.702

Note. These are the results obtained for each of the models with the test data set.

We also used a confusion matrix, and it can be described as a visual representation of the outcomes obtained from the prediction of any classification problem and in each class, the number of correct and wrong predictions are summarized with statistical values (Kokab et al., 2022). The confusion matrices obtained for each model evaluated considering the test data set are through Figure 5 to Figure 10.

Figure 5

Confusion matrix BERT

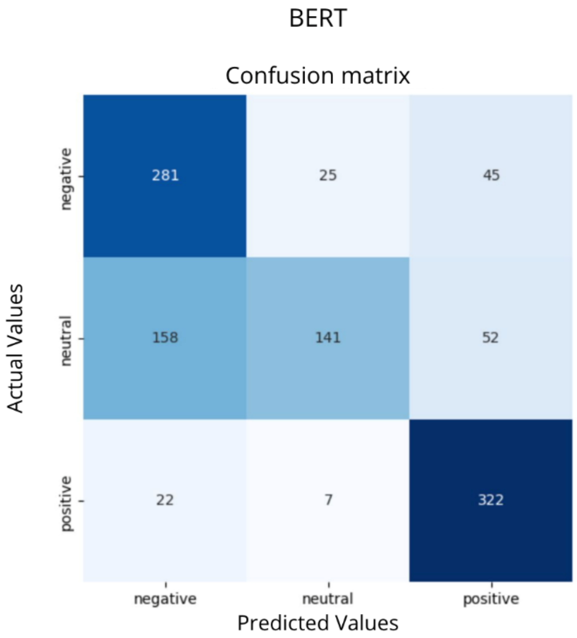


Figure 6

Confusion matrix DistilBERT

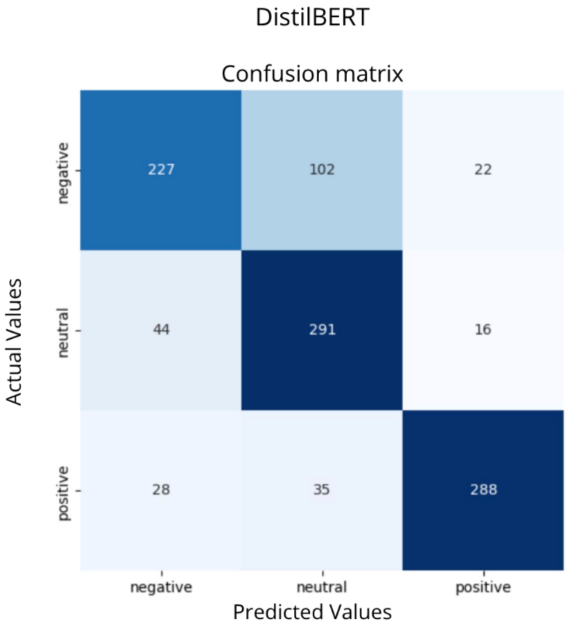


Figure 7

Confusion matrix RoBERTa

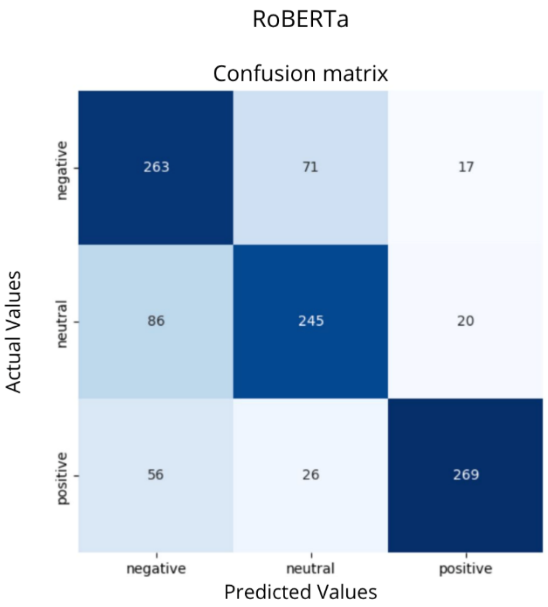


Figure 8

Confusion matrix ChatGPT-4

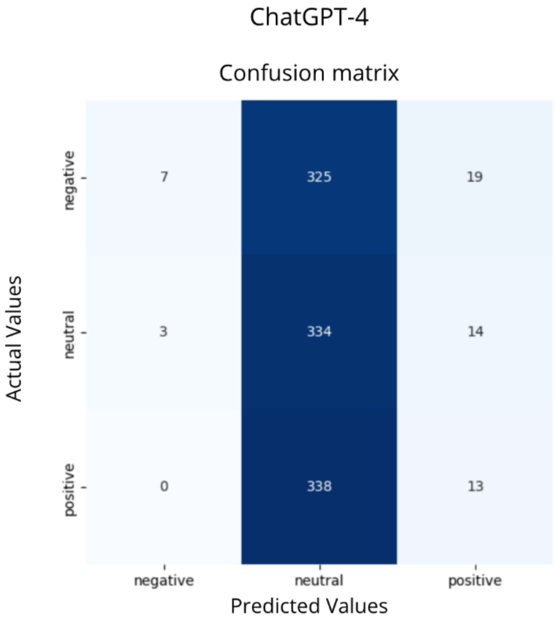


Figure 9

Confusion matrix Google Bard

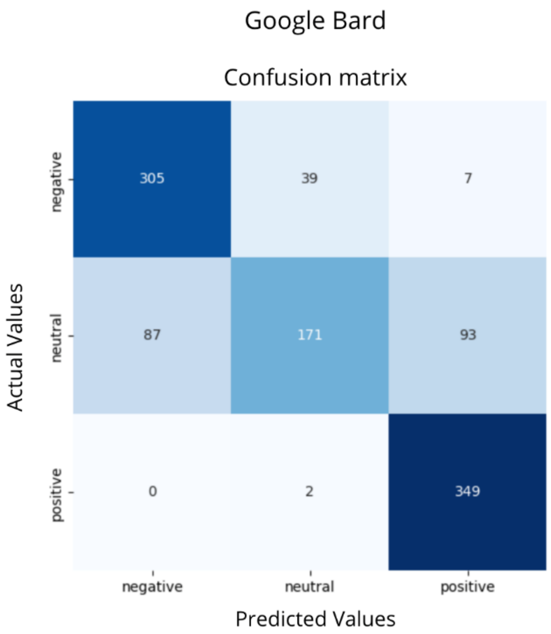
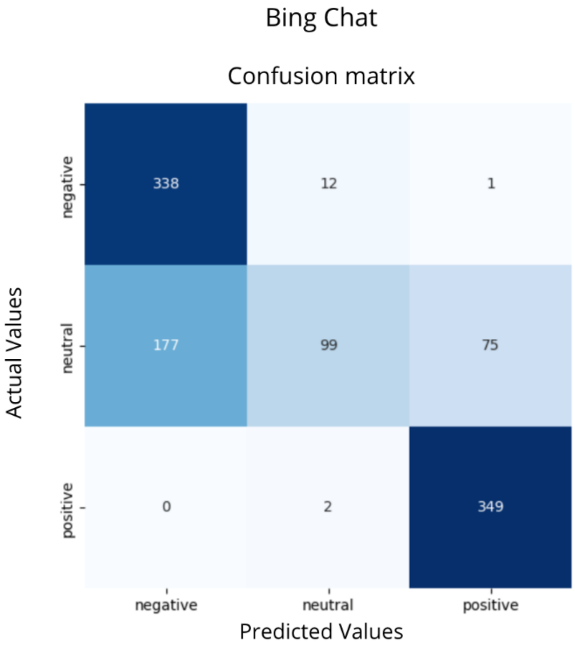


Figure 10

Confusion matrix Bing Chat



As can be seen in the table of accuracies and F1-Scores, the best models were DistilBERT and Google Bard chat, both for the pre-trained models and for the artificial intelligence chats, respectively. It is important to highlight the results from ChatGPT as it has the lowest metrics for performance even though the latest version (GPT-4) was used. On the other hand, to define which of the two mentioned models is better, their advantages and disadvantages were analyzed and placed in comparative Table 4.

Table 4

Key points of comparison between models with better performance

DistilBERT	Google Bard
All comments are entered at the same time.	Comments must be entered one by one to avoid errors.
The training time was slightly extended.	It failed when continuous queries were performed.
A maximum number of tokens was defined.	There is no token limit.
Training in a specific domain.	Multi-domain training.
Manual update.	Automatic update.

2.6. Deployment and control

This phase should focus on the application of the results obtained and how these data should be implemented depending on the complexity required, which can range from delivering a report to having a predictive model in real time (Saltz, 2021). Unfortunately, a real implementation of the chosen model DistilBERT could not be done at USFQ for several reasons:

- The size of the dataset was inadequate for conducting sentiment analysis effectively, as indicated by current literature standards. The inability to obtain the complete database was attributed to concerns regarding confidentiality.
- Permissions to implement the model being students are limited.
- Unclassified data is required to implement the model.

However, this study can serve as a guide for implementing sentiment analysis using DistilBERT, a high-performing model that aligns well with the initially set objectives.

CONCLUSIONS

Sentiment analysis, a versatile tool, has found applications in various fields, including higher education as demonstrated in several academic studies, the paper Pooja and Bhalla (2022) being a review of various research findings in this domain. This study proposes the use of sentiment analysis for evaluating university students' comments about their instructors. The objective was to gather actionable feedback, thereby enabling the University of San Francisco de Quito (USFQ) to enhance its educational quality further. By leveraging machine learning models and AI chat systems, instructors and administrative staff can efficiently analyze student sentiments at the end of a course. This technological approach promises substantial time and resource savings, streamlining the student evaluation process which traditionally relies on manual methods.

Regarding the chosen machine learning models, a variety of options exist for performing different NLP tasks. Specifically, transformer models are notably effective for sentiment analysis due to their ability to contextualize and more deeply understand units of analysis, such as sentences. The three selected transformer models – BERT, RoBERTa, and DistilBERT – have demonstrated their capabilities in sentiment analysis across various domains, including education, as well as in analyzing tweets, movie reviews, and product feedback. Moreover, these models benefit from pre-training on extensive databases, which not only enhances their accuracy but also significantly reduces the required training time and data volume needed for effective deployment. Employing transformer models enables educational institutions to interpret student feedback, identify sentiment trends and recurring themes in comments, and take informed steps to improve teaching quality more accurately. Among these, DistilBERT stands out as the optimal choice, offering a balance of high-performance metrics and efficiency in the classification process.

In addition, a comparative study of robust machine learning models with artificial intelligence chat systems was conducted. In which it was found that, despite requiring more effort and resources, machine learning models are recognized for yielding positive results when implemented correctly. On the other hand, AI chat technologies, such as ChatGPT-4, Google Bard and Bing Chat, have been applied in NLP tasks such as sentiment analysis in other studies. They excel in user interaction, providing simplified and fast responses. Currently, these technologies are advancing rapidly because of this interaction and are being widely applied in tasks that reduce operation times and user effort. Despite their versatility, they may face certain limitations when handling large data sets. Its continued development, however, could overcome these limitations and make it the most effective tool for sentiment analysis in the near future.

Discussing the methodology employed, the CRISP-DM encompasses six crucial stages, each contributing to cleaner, more precise, and accurate results. A pivotal step within this methodology is data preprocessing, especially critical for model training in sentiment analysis. This is because student comments often contain words, signs, and expressions that do not contribute value to the models and thus need to be filtered out. Effective data preprocessing is essential for achieving high accuracy in sentiment analysis, as it ensures the models are trained on relevant, quality data, significantly enhancing their predictive capabilities.

RECOMMENDATIONS

It is recommended to develop a user-friendly interface to achieve correct use of the model. This would allow a natural and simple application of the model, minimizing errors and confusion. Based on this recommendation, it is important to perform maintenance to the model, in this case being a higher-level academic institution, which operates seasonally, the maintenance should follow the calendar of that institution (Klaise, 2020), within the context of this study, it would be the second academic semester.

LIMITATIONS

For the development and training of the different models it is necessary to have computers with high computational power, the limited number of these was a great limitation since it made the development of the project difficult and lengthy. Similarly, for the training of the models it is always advisable to have as much data as possible, for confidentiality reasons the amount of data was limited. Finally, the artificial intelligence chats had some restrictions that increased the time to classify the comments, each of these had a particular characteristic to deliver the results, so for some of them the process was much longer than for others.

BIBLIOGRAPHIC REFERENCES

Acheampong, F.A., Nunoo-Mensah, H. & Chen, W. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif Intell Rev 54, 5789–5829 (2021). https://doi.org/10.1007/s10462-021-09958-2

Ainapure, B. S., Pise, R. N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M. S., & Bizon, N. (2023). Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. Sustainability, 15(3), 2573. https://doi.org/10.3390/su15032573

Ait, A., Izquierdo, J. L. C., & Cabot, J. (2023). HFCommunity: A Tool to Analyze the Hugging Face Hub Community. Proceedings - 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2023, 728–732. https://doi.org/10.1109/SANER56733.2023.00080

Alshamsi, A., Bayari, R., & Salloum, S. (2020). Sentiment analysis in English texts. *Advances in Science, Technology and Engineering Systems Journal*, *5*(6).

Amirita Dewani, Mohsin Ali Memon, & Bhatti, S. (2021). Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00550-7

Aydin, Ö. (2023). Google Bard Generated Literature Review: Metaverse. Journal of AI, 1(7), 1–14. https://dergipark.org.tr/en/download/article-file/3195617

Bakare, A. M., Anbananthen, K. S. M., Muthaiyah, S., Krishnan, J., & Kannan, S. (2023). Punctuation Restoration with Transformer Model on Social Media Data. Applied Sciences (Switzerland), 13(3). https://doi.org/10.3390/app13031685

Caramancion, K. M. (2023). News verifiers showdown: A comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. arXiv preprint arXiv:2306.17176.

Carsten, B. & Eke, D. (2024). The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. International Journal of Information Management, 74, 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999, March). The CRISP-DM user guide. In *4th CRISP-DM SIG Workshop in Brussels in March*(Vol. 1999). sn.

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. Computers and Education: Artificial Intelligence, 1, 100002. https://doi.org/10.1016/j.caeai.2020.100002

Christoph Schröer, Kruse, F., & Jorge Marx Gómez. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Clayson, D. E., & Haley, D. A. (2011). Are students telling us the truth? A critical look at the student evaluation of teaching. *Marketing Education Review*, *21*(2), 101-112.

Cook, C., Jones, A., & Arwa Al-Twal. (2021). Validity and fairness of utilising student evaluation of teaching (SET) as a primary performance measure. Journal of Further and Higher Education, 46(2), 172–184. https://doi.org/10.1080/0309877x.2021.1895093

Dake, D. K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. Education and Information Technologies, 28(4), 4629–4647. https://doi.org/10.1007/s10639-022-11349-1

Delgado Mosquera, A. D. (2012). *Responsabilidad social universitaria enfoque de proyección social en: Universidad San Francisco de Quito y en la Pontificia Universidad Católica del Ecuador* (Bachelor's thesis, Quito, 2012.)

Delobelle, P., Winters, T., & Berendt, B. (2020). Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dodeen, H. (2013). College Students' Evaluation of Effective Teaching: Developing an Instrument and Assessing Its Psychometric Properties. Research in Higher Education Journal, 21. https://eric.ed.gov/?id=EJ1064686

Duong, H.-T., & Tram-Anh Nguyen-Thi. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. Computational Social Networks, 8(1). https://doi.org/10.1186/s40649-020-00080-x

Gan, R. K., Ogbodo, J. C., Wee, Y. Z., Gan, A. Z., & González, P. A. (2023). Performance of Google bard and ChatGPT in mass casualty incidents triage. *The American Journal of Emergency Medicine*.

Gheewala, S., Xu, S., Yeom, S., & Maqsood, S. (2024). Exploiting Deep Transformer Models in Textual Review Based Recommender Systems. Expert Systems with Applications, 235, 121120–121120. https://doi.org/10.1016/j.eswa.2023.121120

Ghorbanali, A., Sohrabi, M.K. A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis. Artif Intell Rev (2023). https://doi.org/10.1007/s10462-023-10555-8

Ghosh, A., Bibhas Chandra Dhara, Pero, C., & Umer, S. (2023). A multimodal sentiment analysis system for recognizing person aggressiveness in pain based on textual and visual information. Journal of Ambient Intelligence and Humanized Computing, 14(4), 4489–4501. https://doi.org/10.1007/s12652-023-04567-z

Giang, N. T. P., Dien, T. T., & Khoa, T. T. M. (2020). Sentiment Analysis for University Students' Feedback. Advances in Intelligent Systems and Computing, 1130 AISC, 55–66. https://doi.org/10.1007/978-3-030-39442-4_5

Gottipati, S., Shankararaman, V. & Lin, J.R. (2018). Text analytics approach to extract course improvement suggestions from students' feedback. RPTEL 13, 6. https://doi.org/10.1186/s41039-018-0073-0

Gupta, V., Velliyur Viswesh, Cone, C., & Unni, E. J. (2020). Qualitative Analysis of the Impact of Changes to the Student Evaluation of Teaching Process. The American Journal of Pharmaceutical Education, 84(1), 7110–7110. https://doi.org/10.5688/ajpe7110

Haakman, M., Cruz, L., Hennie Huijgens, & Arie van Deursen. (2021). AI lifecycle models need to be revised. Empirical Software Engineering, 26(5). https://doi.org/10.1007/s10664-021-09993-1

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. Procedia CIRP, 79, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Hugging Face. (2018).

Huynh, L. M., Bonebrake, B. T., Schultis, K., Quach, A., & Deibert, C. M. (2023). Google Bard Artificial Intelligence vs the 2022 Self-Assessment Study Program for Urology. *Urology Practice*, *10*(6), 553-555.

Kalra, V., & Aggarwal, R. (2017). Importance of Text Data Preprocessing & Implementation in RapidMiner. *ICITKM*, *14*, 71-75.

Kaminska, O., Cornelis, C., & Hoste, V. (2023). Fuzzy Rough Nearest Neighbour Methods for Aspect-Based Sentiment Analysis. Electronics (Switzerland), 12(5). https://doi.org/10.3390/electronics12051088

Kayvan Tirdad, Alex Dela Cruz, Sadeghian, A., & Cusimano, M. D. (2021). A deep neural network approach for sentiment analysis of medically related texts: an analysis of tweets related to concussions in sports. Brain Informatics, 8(1). https://doi.org/10.1186/s40708-021-00134-4

Klaise, J., Van Looveren, A., Cox, C., Vacanti, G., & Coca, A. (2020). Monitoring and explainability of models in production. *arXiv preprint arXiv:2007.06299*.

Kokab, S. T., Asghar, S., & Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. Array, 14, 100157.

Kotei, E., & Thirunavukarasu, R. (2023). A Systematic Review of Transformer-Based PreTrained Language Models through Self-Supervised Learning. In Information (Switzerland) (Vol. 14, Issue 3). MDPI. https://doi.org/10.3390/info14030187

Krouska, A., Troussas, C., & Virvou, M. (2016, December 14). The effect of preprocessing techniques on Twitter sentiment analysis. IISA 2016 - 7th International Conference on Information, Intelligence, Systems and Applications. https://doi.org/10.1109/IISA.2016.7785373

Lazrig, I., & Humpherys, S. L. (2022). Using Machine Learning Sentiment Analysis to Evaluate Learning Impact. Information Systems Education Journal, 20(1), 13-21.

Laifa, M., Giglou, R. I., & Akhrouf, S. (2023). Blended Learning in Algeria: Assessing Students' Satisfaction and Future Preferences Using SEM and Sentiment Analysis. Innovative Higher Education. https://doi.org/10.1007/s10755-023-09658-5

Liu, B. (2015). Sentiment analysis: mining opinions, sentiments, and emotions. The Cambridge University Press.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1-35.

Mabunda, J. G. K., Jadhav, A., & Ajoodha, R. (2021). Sentiment analysis of student textual feedback to improve teaching. Interdisciplinary Research in Technology and Management, 643-651.

Malik, Назарова, А. Г., Jamjoom, M., & Ignatov, D. I. (2023). Multilingual Hope Speech Detection: A Robust Framework Using Transfer Learning of Fine-tuning Roberta Model. Journal of King Saud University - Computer and Information Sciences, 35(8), 101736–101736. https://doi.org/10.1016/j.jksuci.2023.101736

Manias, G., Mavrogiorgou, A., Kiourtis, A., Chrysostomos, S. & Dimosthenis, K. (2023). Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Comput & Applic 35, 21415–21431. https://doi.org/10.1007/s00521-023-08629-3

Martínez, L. A., Sandoval, A. L., & García, L. J. (2023). Analysis of Digital Information in Storage Devices Using Supervised and Unsupervised Natural Language Processing Techniques. Future Internet, 15(5). https://doi.org/10.3390/fi15050155

Mathew, L., & Bindu, V. R. (2022). Efficient Transformer Based Sentiment Classification Models. Informatica (Slovenia), 46(8), 175–184. https://doi.org/10.31449/inf.v46i8.4332

Miranda, C. H., Sanchez-Torres, G., & Salcedo, D. (2023). Exploring the Evolution of Sentiment in Spanish Pandemic Tweets: A Data Analysis Based on a Fine-Tuned BERT Architecture. Data, 8(6). https://doi.org/10.3390/data8060096

Mohd, Hijazi, A., Ervin Gubin Moung, Puteri N. E. Nohuddin, Chua, S., & Coenen, F. (2023). Social Media Sentiment Analysis and Opinion Mining in Public Security: Taxonomy, Trend Analysis, Issues and Future Directions. Journal of King Saud University - Computer and Information Sciences, 101776–101776. https://doi.org/10.1016/j.jksuci.2023.101776

Moons, P., & Van Bulck, L. (2023). Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. European Journal of Cardiovascular Nursing, zvad087.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology.

Moussa, A., Fournier, S., Mahmoudi, K., Espinasse, B., & Sami Faïz. (2022). Mixing Static Word Embeddings and RoBERTa for Spatial Role Labeling. Procedia Computer Science, 207, 2950–2957. https://doi.org/10.1016/j.procs.2022.09.353

Nelson, S., Walsh, C., Olsen, C., McLaughlin, A., LeGrand, J., Schutz, N., & Lasko, T. (2020). Demystifying artificial intelligence in pharmacy. American journal of health-system pharmacy: AJHP : official journal of the American Society of Health-System Pharmacists. https://doi.org/10.1093/ajhp/zxaa218.

Ngoc, T. V., Thi, M. N., & Thi, H. N. (2021). Sentiment Analysis of Students' Reviews on Online Courses: A Transfer Learning Method. In Proceedings of the International Conference on Industrial Engineering and Operations Management.

Nguyen, P., Trương, H., Nguyen, P., Bruneau, P., Cao, L., & Wang, J. (2023). Evaluation of Google Bard on Vietnamese High School Biology Examination. Researchgate. Net.

Parnami, A., & Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning. arXiv preprint arXiv:2203.04291.

Pooja, & Bhalla, R. (2022). A review paper on the role of sentiment analysis in quality education. SN Computer Science, 3(6), 469. https://doi.org/10.1007/s42979-022-01366-9

Pustikayasa, I. M., Purnawati, N. W., & Arsana, I. N. A. (2023). Bing Chat AI for Learning Supplement. *International Proceeding on Religion, Culture, Law, Education, And Hindu Studies*, *1*, 11-17.

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2023). Improving language understanding by generative pre-training. 2018. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

Rahaman, M. S., Ahsan, M. M., Anjum, N., Rahman, M. M., & Rahman, M. N. (2023). The AI race is on! Google's Bard and OpenAI's ChatGPT head to head: an opinion article. *Mizanur and Rahman, Md Nafizur, The AI Race is on*

Rani, D., Kumar, R., & Chauhan, N. (2022). Study and Comparision of Vectorization Techniques Used in Text Classification. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt54827.2022.9984608

Riadh Meghatria, Chiraz Latiri, & Nader, F. (2020). Event Nugget Detection using Pre-trained Language Models. Procedia Computer Science, 176, 320–329. https://doi.org/10.1016/j.procs.2020.08.034

Rupp, M. T. (2023). A Critical Look at Student Evaluations of Teaching. The American Journal of Pharmaceutical Education, 100136–100136. https://doi.org/10.1016/j.ajpe.2023.100136

Salazar, C., Montoya-Múnera, E. & Aguilar, J. Sentiment analysis in learning resources. J. Comput. Educ. 10, 637–662 (2023). https://doi.org/10.1007/s40692-022-00237-9

Saltz, J. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 2337-2344, doi: 10.1109/BigData52589.2021.9671634.

Santisteban, L., & Egues, A. L. (2022). How student evaluations of teaching contribute to hindrance of faculty diversity? Teaching and Learning in Nursing, 17(4), 455–459. https://doi.org/10.1016/j.teln.2022.04.007

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3). https://doi.org/10.1007/s42979-021-00592-x

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. Natural Language Processing Journal, 2, 100003. https://doi.org/10.1016/j.nlp.2022.100003

Siddharth, & R Aarthi. (2023). Blended multi-class text to image synthesis GANs with RoBerTa and Mask R-CNN. Procedia Computer Science, 218, 845–857. https://doi.org/10.1016/j.procs.2023.01.065

Sproule, R. (2000). Student evaluation of teaching: Methodological critique. *Education Policy Analysis Archives*, *8*, 50-50.

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. http://arxiv.org/abs/1903.09588

Tabinda Kokab, S., Asghar, S., & Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. Array, 14. https://doi.org/10.1016/j.array.2022.100157

Tang, T., Tang, X., & Yuan, T. (2020). Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text. IEEE Access, 8, 193248–193256. https://doi.org/10.1109/ACCESS.2020.3030468

Tian, L., Lai, C., & Moore, J. D. (2018). Polarity and intensity: the two aspects of sentiment analysis. arXiv preprint arXiv:1807.01466. http://dx.doi.org/10.18653/v1/W18-3306

Vargas-Madriz, & Nocente, N. (2023). Exploring students' willingness to provide feedback: A mixed methods research on end-of-term student evaluations of teaching. Social Sciences & Humanities Open, 8(1), 100525–100525. https://doi.org/10.1016/j.ssaho.2023.100525

Vásquez, J., Gómez-Adorno, H., & Bel-Enguix, G. (2021). Bert-based Approach for Sentiment Analysis of Spanish Reviews from TripAdvisor. https://github.com/juanmvsa/Sentiment-Analysis-TripAdvisor-Spanish

Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55, 5731–5780 (2022). https://doi.org/10.1007/s10462-022-10144-1

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Xuan-Quy, D., Ngoc-Bich, L., Xuan-Dung, P., Bac-Bien, N., & The-Duy, V. (2023). Evaluation of ChatGPT and Microsoft Bing AI Chat Performances on Physics Exams of Vietnamese National High School Graduation Examination. arXiv preprint arXiv:2306.04538.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

APPENDIX A: GITHUB LINK WITH THE CODES OF THE DATA PREPROCESSING, TRAINED MODELS, AND ANALYSIS OF THE RESULTS OF THE AI CHATS

https://github.com/joseval152001/Sentiment-analysis-with-transformers