Sentiment Analysis in the Feedback of Peer Evaluation Activities

René Elizalde-Solano[1], Ma.Carmen Cabrera-Loayza[1,2], Elzabeth Cadme[1,2] and Nelson Piedra[1,2]

[1] *Universidad Técnica Particular de Loja, San Cayetano 1101608, Loja, Ecuador*
[2] *Universidad Politécnica de Madrid, Boadilla del Monte 28660, Madrid, España*

Abstract

Sentiment analysis is a technique used more frequently in the educational field. For the present work, the analysis and classification of the feedback comments issued by the students in the peer evaluation activities has been taken as the main application approach. Determining the polarity of these comments can help the teacher to identify characteristics and patterns in the criteria issued by the students to enrich the teaching-learning process. The present work aims to determine the polarity of feelings of the feedback comments of the peer evaluation activities planned as challenges within the courses offered by the Open Campus initiative. To do this, experimentation is carried out in three training scenarios and tests of the classification model using the corpus of tweets written in Spanish TASS and a corpus of comments extracted from the learning platform, manually classified by experts. Among the main results, it is observed that many students give feedback that is useful, be it positive or negative. However, there is a significant percentage of comments that are perceived as unjustified or incomprehensible, and this is observed in the number of comments classified as neutral and without polarity.

Keywords

Sentiment Analysis, Peer Assessments, Open Campus, Feedback, Open Online Courses, Open Education

1. Introduction

Currently, the design and planning of online courses, a number of evaluation and training activities are defined. It is intended that students acquire, beyond professional competencies, some soft skills within the teaching-learning process. Within the Open Campus initiative, the collaborative work of students is encouraged to create learning communities guided by a teacher and enriched by the participants. One of the main evaluation proposed activities in each course offered is called "challenge". Challenges are peer review activities that allow students to review, evaluate, and provide feedback on the work of their peers. This guarantees student is the main actor of the assessment process carried out, also acquired skills such as collaborative work, co-construction of knowledge, reflection, and critical assessment [1].

Students' general comments about the evaluation they have made of assigned work. Generally, these feedbacks or opinions are not mandatory, therefore are not considered in this analysis, and only the grades given are considered. The main objective of this work is to determine the polarity of feelings in the feedback comments of the peer evaluation activities posed as challenges within the courses offered by the Open Campus initiative. The experimentation in three scenarios is approached for the training and testing of the classification model. In the first scenario, TASS 2019 corpus is used [2]. In the second scenario, a manually classified corpus of comments from the Open Campus platform is used. For the third scenario, the model is trained with a mixture of the data mentioned above Finally, it should be mentioned that the comments are in Spanish and that the Linear Support Vector Classification algorithm is applied for each scenario [3].

CISETC 2021: International Congress on Educational and Technology in Sciences, November 16-18, 2021, Chiclayo, Peru

EMAIL: rrelizalde@utpl.edu.ec (R. Elizalde-Solano); mccabrerax@utpl.edu.ec (MC. Cabrera-Loayza); iecadme@utpl.edu.ec (E. Cadme); nopiedra@utpl.edu.ec (N. Piedra)
ORCID: 0000-0002-9534-8450 (R. Elizalde-Solano); 0000-0002-7664-5206 (MC. Cabrera-Loayza); 0000-0002-5554-0560 (E. Cadme); 0000-0003-1067-8707 (N. Piedra)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Sentiment Analysis

2.1. Feedback - Peer reviews

Some actors have argued peer evaluation is a particularly useful practice of training activities because students need to develop their own evaluations skills to better recognize quality, understand evaluation criteria, and self-evaluate their own work [4]. Some actors have argued peer evaluation is a particularly useful practice of training activities because students need to develop their own evaluations skills to better recognize quality, understand evaluation criteria, and self-evaluate their own work [4]. This includes those students who can benefit both from receiving feedback from their peers and from building feedback on the work of others, and some research has determined that giving feedback improved writing performance as well as how to receive feedback [5].

In [6] peer assessment is defined as a teaching-learning strategy that allows students to provide peer feedback. Despite the benefits of peer review, it is always an arduous process for any teacher who explores some meaningful information about decision-making [7].

Therefore, it is important to analyze feedback comments given by students with the help of computational techniques, such as machine learning. In order to determine which are the most important aspects that learners consider when evaluating the work of their peers. In addition, through the comments, the perception and understanding of the students about the proposed activity can also be identified [8]. In addition, through the comments, the perception and understanding of the students about the proposed activity can also be identified [8]. Also, patterns are identified in the relationship between the student's opinions, the feedback they give to other students, and how they react to the feedback they receive.

2.1.1. Analysis of feelings in the educational context

Sentiment analysis is a task that focuses on detecting polarity and recognizing the emotion that an individual may feel about a topic, or event. The main goal of sentiment analysis is to find the opinions of users, identify the feelings they express, and then classify their polarity into positive, negative, and neutral categories.

Sentiment analysis systems use Natural Language Processing techniques as well as Machine Learning to discover, retrieve and extract information and opinions from large amounts of textual information [9]. Sentiment analysis and opinion mining are similar. But there is a slight difference, the former refers to finding feeling words and phrases that show emotions, while the latter refers to extracting and analyzing opinions of people for a given entity [8]. Sentiment analysis is a field of research that has grown rapidly in recent years in the context of student comments in learning platform environments [10].

When searching the term “sentiment analysis” in the Scopus database, results in about 19,000 papers at a general level. However, in the educational context, there are around 80 papers and few of them refer to the analysis of the students' comments obtained in the peer evaluation-type activities. In [11] a study on sentiment analysis in the educational context is carried out focuses on detecting the approaches and digital educational resources used in the sentiment analysis, as well as identifying the main benefits of using this analysis in the domain of education. The results show that Naïve Bayes is the most used technique and that the forums in MOOC and social networks are the most used digital education resources to collect the data necessary to carry out the sentiment analysis process.

On the other hand, in [7] a study of several experiments is carried out with a manually labeled dataset to test different combinations of N-grams with inverse document term-frequency frequency (TF-IDF) and classification algorithms. As result, it is obtained that the Support Vector Machine classifier combining 1 gram + 2 grams + TF-IDF considered the best model in Precision, Recall and F-Measure

In the study exposed in [12], it was determined that the students who considered the feedback useful tended to be more receptive when acknowledging their mistakes, while the students who found the feedback less useful tended to be more defensive when expressing that they were confused about the comments, and they disagreed with the statements given. Finally, the study carried out in [13] focuses on determining the inconsistencies that arise in the peer evaluations, between the numerical score and the textual feedback. Experiments carried out with 4 student groups and 2 activity types have determined that the general peer evaluation process is a process with reliable results, which guarantees a valuable approach to ensure the correct functioning of the peer review process.

3. Methodology

The process carried out to analyze the polarity of the comments issued in the peer evaluation activities of the courses on the Open Campus platform is detailed below. First, an ETL process is performed to extract the data set from the comments. Then, a process of cleaning the information is carried out to later apply the classification algorithm and evaluate the performance using the precision metrics, the F-Score measure, and the confusion matrix.

The next task is to find a corpus in Spanish that allows training the classification models for their subsequent application to the set of feedback comments. This task had difficulties since there are not many corpora in Spanish available. For the present work, the corpus generated in the Workshop on Semantic Analysis at SEPLN (TASS) [14] is used, which compiles a set of tweets written in Spanish. In addition, a corpus is also created with the comments of the feedback from the peer evaluations of the Open Campus platform, manually labeled by experts as positive, negative, neutral, and none (none). Finally, the classification models are trained in three scenarios that are detailed in the next section.

3.1. Training and testing phase

Next, the training and test phase is developed in the three proposed scenarios:

3.1.1. Scenario 1

With the TASS corpus, we proceed to extract the necessary data to apply the classification algorithm with the comments of our context. It is important to indicate that some Python programming language libraries are used, such as Pan-das [15], Scikit Learn [16], NLTK [17]. Scikit Learn makes use of the supervised algorithm of Linear Support Vector Classification. The NLTK library uses it to generate a function that allows comments to be tokenized.

Model training. The set of already classified comments used to train the selected model was a total of 7608; each one of them categorized as positive, negative, neutral and none, see Figure 1a.

Before being able to apply the Linear Support Vector Classification algorithm, the CountVectorizer function is used, which allows each comment to be separated into a frequency vector for each word that composes it. When working with information in Spanish, procedures were specified to refine the vectorization process, such as not considering stopword in Spanish, using the SnowballStemmer algorithm to join words based on their root, and through the word_tokenize method of the NLTK library to separate each word into its respective syllables. The result of CountVectorizer is a data frame 5706 rows and 9754 columns.

To generate the classification model, LinearSVC from the Scikit-learn library is used. It is important to highlight that to train the algorithm, the information of the comments is sent, but at the level of numerical vectors, together with the labeling of each expression.

Model test. Once the model has been trained, it starts by separating the information to be used for training and testing; For this, the train_test_split function of the Scikit Learn library was the mechanism that allows having 5706 comments for training and 1902 comments for tests.

As a result of the test phase of the model, there is an accuracy of 71.66% through the accuracy_score metric of Scikit-Learn and 68.45% through the f1_score metric. The confusion matrix after applying the algorithm mentioned is detailed in Figure 1b.

3.1.2. Scenario 2

Scenario 2 looks for a way to create a classified data set from the context of peer reviews of the Open Campus platform.

Model training. From the set of 101559 comments extracted, a data set of 2992 comments are generated randomly. This data set was manually classified by experts to assign polarity according to their criteria. Figure 1c shows the result of manual classification.

This new data set will be used to train the Linear Support Vector Classification algorithm. Before doing so, as indicated in scenario 1, the data set is divided for training 2244 records and test 748 records. Furthermore, the CountVectorizer function is used to vectorize the information set, obtaining a data frame of 2244 rows and 2171 columns. Finally, the classification algorithm LinearSVC is applied.

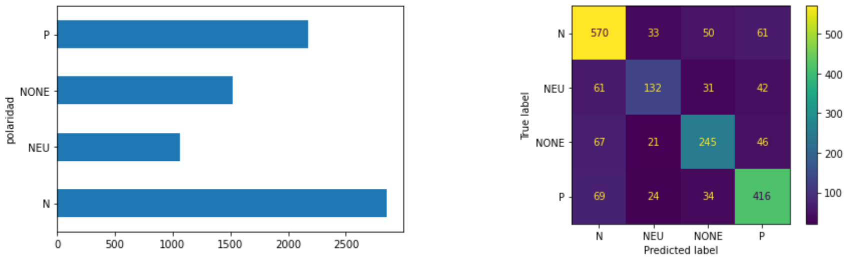
Model test. Once the model has been trained, the model is evaluated with the 748 records. The result of applying the algorithm provides the following data, an accuracy of 73.66% through the Scikit-learn accuracy_score metric and 56.28% through the f1_score metric. The confusion matrix is detailed in Figure 1d.

3.1.3. Scenario 3

For Scenario 3, the research team decides to pool the trained dataset. Use is made of classified information from the TASS and comments manually classified by experts.

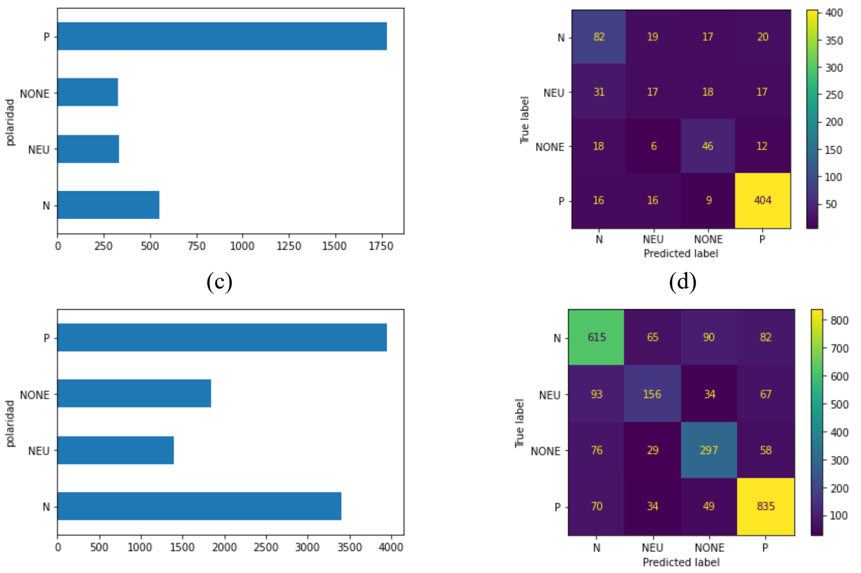
Model training. For the training phase, a data set with 10,600 records is consolidated, classified according to their polarity, as can be seen in Figure 1e. As in scenarios 1 and 2, the data set is generated, for training 7950 data and for testing 2650 data. The information is vectorized through CountVectorizer obtaining a dataframe of 7950 rows and 16316 columns, and the Linear Support Vector Classification algorithm is applied.

Model test. Once the model has been trained, we proceed to evaluate the model with the 748 records. And an accuracy of 70.67% is obtained through the accuracy_score metric of Scikit-Learn and 65.76% through the metric f1_score. Furthermore, the confusion matrix is obtained, see Figure 1f.



(a)

(b)



(c) (d)

(e) (f)

Figure 1: Classified comments and confusion matrix for each scenario: (a) polarity of TASS corpus comments, (b) confusion matrix - scenario 1, (c) polarity of the manually created corpus, (d) confusion matrix - scenario 2, (e) TASS corpus data set and those manually classified from the Open Campus platform, (d) confusion matrix - scenario 3.

3.2. Classification phase

3.2.1. Classification using the scenario 1 model

The process is carried out to determine the polarity of 101,559 comments from the peer reviews of the Open Campus platform and apply the trained model to each of them. The results are as follows see Table 1 and Figure 2a.

Table 1
Comment Rating - Scenario 1

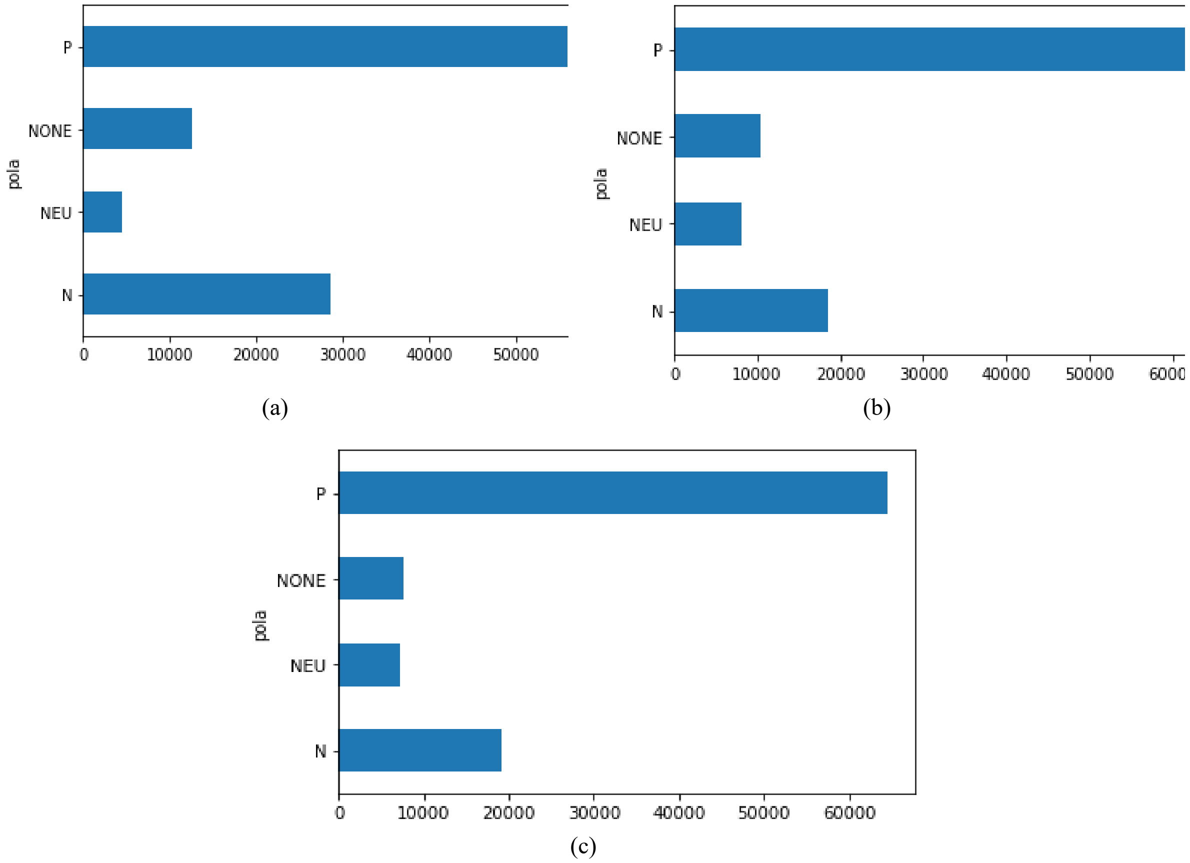
	Feedback Polarity			
	Positive (P)	Negative (N)	Neutral (NEU)	No Polarity (NONE)
Total comments	55975	28537	4469	12578

3.2.2. Classification using the scenario 2 model

The 98567 comments from the peer reviews of the Open Campus platform are classified and the trained model is applied to each of them. The following results were obtained, see Table 2 and Figure 2b.

Table 2
Comment Rating - Scenario 2

	Feedback Polarity			
	Positive (P)	Negative (N)	Neutral (NEU)	No Polarity (NONE)
Total comments	61622	18443	8035	10459



(a)

(b)

(c)

Figure 2: Comments classified based on the trained model in each scenario (a) comments classified with scenario 1, (b) comments classified with scenario 2, (c) comments classified with scenario 3.

3.2.3. Classification using the scenario 3 model

Then, the 98,559 comments from the peer reviews of the Open Campus platform are classified and the trained model is applied to each of them. The following results were obtained, see Table 3 and Figure 2c.

Table 3
Comment Rating - scenario 3

	Feedback Polarity			
	Positive (P)	Negative (N)	Neutral (NEU)	No Polarity (NONE)
Total comments	64235	19247	7089	7901

4. Results and discussion

In this research, scenarios were created to analyze the set of comments expressed by the participants of the Open Campus platform courses. Table 4 shows the polarity obtained from the classified feedback comments with the trained models in each scenario.

Table 4
Classification of comments by stage

Scenarios	NcT	Feedback Polarity							
		Positive		Negative		Neutral		No Polarity	
		NC	%	NC	%	NC	%	NC	%
Scenario 1	101559	55975	55.11	28537	28.08	4469	4.4	12578	12.3
Scenario 2	98559	61622	62.52	18443	18.71	8035	8.15	10459	10.61
Scenario 3	98559	64235	65.17	19247	19.16	7089	7.19	7901	8.01

As can be seen in Table 4, for each scenario, a similar number of total comments (NcT) is classified. Based on this data set and the previously trained classification model, it is observed that the trend in the types of polarity in the three scenarios is equivalent since there is a greater polarity of *positive* comments from the participants. In order of polarity, *negative* comments are the second most frequent. However, it is observed that comments classified as *non-polar* have a higher number of occurrences than comments classified as *neutral*. This is because many comments do not contribute to feedback or cannot be framed in context. Furthermore, it is observed that when comparing scenarios 1 and 2, there is a considerable difference in the polarity classification percentage. This is because for scenario 1 only comments from the TASS corpus are used. And for scenario 2, the platform's own comments classified manually are used. With this, it is determined that while the model is trained with data closer to the context, the classification will be more reliable within the types of polarity proposed.

With respect to scenario 3, an improvement in the classification of positives and negatives is observed. This is attributed to the fact that there is a larger number of training data than the previous scenarios, and that the TASS data set and the domain's own data set are involved for training. Even though the domain dataset is smaller in this scenario, the classification is more accurate. According to Figure 3, taking with reference the variation in the number of positive comments that the model generates, it is evident that scenario three has the highest number of positive comments. It is emphasized that said scenario has the following advantages: a greater number of trained data and information related to the context of the comments of the Open Campus platform.

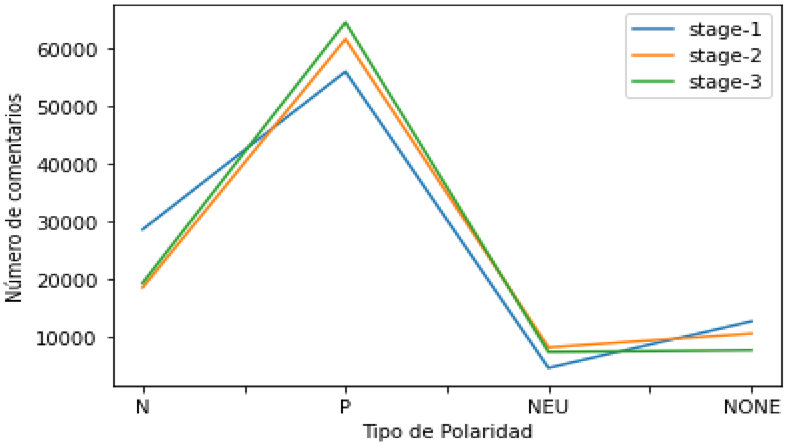


Figure 3. Results of the classification and polarity of comments considering the three proposed scenarios.

5. Conclusions

In the present work, it is determined that the information within the feedback comments of the peer evaluation activities has great potential for both teachers and participants. For teachers this information can give a vision of how students perceive the activity and the contributions of their peers from a qualitative point of view. And from the students' side, evaluating the activities of their classmates allows them to have a better understanding of the subject of study and develop soft skills such as critical thinking, co-evaluation, and collaborative work. Furthermore, analyzing the results obtained, it is identified that many students give feedback that is useful, whether it is positive or negative. However, there is an important percentage of comments in the feedback that is perceived as unjustified or incomprehensible, and this is observed in the number of comments classified as neutral and without polarity. Finally, it is stated that the more context data is used in the training phase, the more remarkable the accuracy in the classification. In addition, with the present work it has been observed that there is no corpus in Spanish related to the educational field. This research is a contribution to future works that require a corpus of comments in Spanish for feedback.

6. Acknowledgements

This research is supported by the Knowledge-Based Systems Research Group of the UTPL and the Open Campus initiative, Loja - Ecuador

7. References

[1] N. Osheroff, W. B. Cutrer, C. C. Pettepher, R. H. Carnahan, and E. C. Bird, “Using Small Case-Based Learning Groups as a Setting for Teaching Medical Students How to Provide and Receive Peer Feedback,” Med. Sci. Educ., vol. 27, no. 4, pp. 759–765. (2017). doi: 10.1007/s40670-017-0461-x

[2] Martínez-Cámara, E., García-Cumbreras, M.A., Villena-Román, J., García-Morera, J. TASS 2015 - La evolución de los sistemas mineros de opinión españoles. Procesamiento del Lenguaje Natural, 56. (2020).

[3] Esparza, G. G., de-Luna, A., Zezzatti, A. O., Hernandez, A., Ponce, J., Álvarez, M., ... & de Jesus Nava, J. A sentiment analysis model to analyze students reviews of teacher performance using support vector machines. In International Symposium on Distributed Computing and Artificial Intelligence (pp. 157-164). Springer, Cham. (2017). doi:10.1007/978-3-319-62410-5_19

[4] Sadler, D. R. Beyond feedback: Developing student capability in complex appraisal. Assessment & Evaluation in Higher Education, 35(5), 535–550. (2010). doi:10.1080/02602930903541015.

[5] Lundstrom, K., & Baker Smemoe, W. To give is better than to receive: The benefits of peer review to the reviewer’s own writing. Journal of Second Language Writing, 18, 30–43. 2009. doi:10.1016/j.jslw.2008.06.002.

[6] H. Shang, “An exploration of asynchronous and synchronous feedback modes in EFL writing,” J. Comput. High. Educ., vol. 29, no. 3, pp. 496–513, 2017.]

[7] Ortega, M. P., Mendoza, L. B., Hormaza, J. M., & Soto, S. V. Accuracy'Measures of Sentiment Analysis Algorithms for Spanish Corpus generated in Peer Assessment. In Proceedings of the 6th International Conference on Engineering & MIS 2020. pp. 1-7. (2020). doi:10.1145/3410352.3410838

[8] Misiejuk, K., Wasson, B. and Egelandsdal, K. Using learning analytics to understand student perceptions of peer feedback. Computers in human behavior, 117, p.106658. (2021). doi: 10.1016/j.chb.2020.106658

[9] Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intell. Syst. (2013). doi:10.1109/MIS.2013.30

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

[11] Mite-Baidal, K.; Delgado-Vera, C.; Solís-Avilés, E.; Espinoza, A.H.; Ortiz-Zambrano, J.; Varela-Tapia, E. Sentiment Analysis in Education Domain: A Systematic Literature Review. In International Conference on Technologies and Innovation; Valencia-García, R., Alcaraz-Mármol, G., Del Cioppo-Morstadt, J., Vera-Lucio, N., Bucaram-Leverone, M., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, pp. 285–297. (2018). doi:10.1007/978-3-030-00940-3_21

[12] Zong, Z., Schunn, C. D., & Wang, Y. (2021). What aspects of online peer feedback robustly predict growth in students' task performance? *Computers in Human Behavior*, 106924. doi:10.1016/j.chb.2021.106924

[13] Rico-Juan, J. R., Gallego, A. J., & Calvo-Zaragoza, J. (2019). Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine learning. Computers & Education, 140. (2019). doi:10.1016/j.compedu.2019.103609

[14] TASS: Workshop on Semantic Analysis at SEPLN. http://tass.sepln.org/

[15] Pandas: https://pandas.pydata.org/

[16] Scikit-Learn: https://scikit-learn.org/

[17] NLTK Library: https://www.nltk.org/