Automatic classification of subjective attributes from student messages in virtual learning environments

Diego Rossi[1], Victor Ströele[1], Jairo Souza[1], Fernanda Campos[1]

[1]Programa de Pós-graduação em Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF) - MG - Brasil

diego.rossi@ice.ufjf.br, victor.stroele@ufjf.edu.br,

jairo.souza@ice.ufjf.br, fernanda.campos.ufjf@gmail.com

Abstract. *Accompanying students in virtual learning environments to identify those who need help is a difficult and time-consuming task. Identifying subjective attributes that recognize students' feelings can help teachers and tutors in pedagogical interventions. The interaction must motivate and keep students engaged. This article proposes an architecture capable of automatically carrying out some pedagogical intervention. The architecture uses automatic textual classification models to detect multiple attributes in post messages, such as Sentiment, Post Type, Urgency, and Confusion. The main goal is to predict learning problems and act to minimize their impacts. The evaluation was performed with data from the Stanford MOOCPosts Dataset to verify if the models allow the automatic classification of subjective attributes. Our results show that the proposal outperforms other approaches in this dataset.*

1. Introduction

The popularization and advances in Information Technology (IT) have made it increasingly part of our daily lives. Consequently, due to its relevance, it is widely used in the educational environment [Panigrahi et al. 2020], supporting innovative pedagogical practices, and generating new learning spaces [Moreno-Guerrero et al. 2020]. Furthermore, IT allows people to study at distance modalities, which also democratizes access to education since prices are more attractive due to a leaner structure [Palvia et al. 2018]. Another point contributing to increased distance learning is study schedules, which make it more flexible and enable the student to have a full-time job.

According to [Fandiño and Velandia 2020] and [Cuevas et al. 2018], one main factor that positively affects learning is motivation. Certainly, tutors play a very important role in this process, where they are responsible for interacting with the students to motivate and keep them engaged. To make tutoring more agile and efficient, even automating some tasks, it is necessary to identify which students need specific help. Furthermore, to select the students who will undergo an intervention, it is necessary to identify implicit characteristics in their messages to provide accurate pedagogical intervention.

These challenges demand that tutors accompany students agilely, providing a communication environment capable of answering questions and motivating students [Moreno-Guerrero et al. 2020]. So, we believe that assisted education may contribute to the automatization process of student tutoring. In this context, there are two major challenges: the first is sending recommendations and academic topics to students; The second is to predict learning problems and act to minimize their impacts [Toti et al. 2020].

DOI: 10.5753/sbie.2022.224725

Discussion forums are among the most popular interaction tools offered by Learning Management Systems (LMS), often used by students to create a sense of belonging and better understand course topics [Capuano and Caballé 2015]. According to [Capuano and Caballé 2019] it is possible, through natural language processing (NLP) and predictive models, to detect multiple attributes in post messages, such as Sentiment, Post Type, Urgency, and Confusion. These attributes underlie the choice of pedagogical intervention action. Thus, the semantic detection of forum posts offers implicit information, fundamental for more careful analysis to evaluate the student knowledge, consequently collaborating to moderate and plan interventions. However, students trying to clarify concepts through these forums may not get the attention they need, and the lack of responsiveness often favors dropout [Capuano and Caballé 2019].

Some computer systems seek to fill this gap through automatic student interactions, such as Conversational Agents (CA), Recommender Systems (RS), Chatbots, etc. Usually, these systems use Artificial Intelligence (AI) approaches operating based on a well-defined set of rules that shape their behavior when interacting with humans [Demetriadis et al. 2021]. Researchers have already applied these computer systems to accomplish various educational goals such as tutoring, question-answering, language learning practice, and the development of metacognitive skills [Khanal et al. 2020].

The objective of this work is to predict learning problems and act to minimize their impacts. Knowing students' educational moments is necessary to identify those who need help. The automatic textual classification is proposed to identify subjective attributes in messages from students in virtual learning environments. The attributes identified were Sentiment, Confusion, and Urgency, to determine the most appropriate pedagogical intervention for the student.

This article is organized as follow: Section 2 presents related work. Section 3, the proposed architecture, details the main layers that make up the solution workflow. The results achieved are presented in Section 4. In section 5 are the final remarks.

2. Related work

[Capuano and Caballé 2019] use a rich data source to tackle discussion forums in a MOOC environment. They proposed a multi-attribute text categorization tool that automatically detects useful information from forum posts, including intents (question, answer, and opinion), topics, sentiment polarity, and levels of confusion and urgency.

[Bóbó et al. 2019] present the SASys architecture, based on a lexical approach and a polarized frame network. The main goal is to identify the author's sentiment in texts. The semantic orientation of the text is determined by the result of sentiment analysis approaches (Lexical, Machine Learning, or Hybrid). Then, a recommendation system, based on the student's emotional state and learning style, sends motivational messages to mitigate the dropout.

According to [Moreno-Marcos et al. 2018b], e-Learning platforms store a large amount of data from all student interactions. These interactions concern course accesses, video events (when the video was played, speed, etc.), and exercise logs (attempts, scores, tips used, etc.). This data can predict behavior and outcomes, enabling teachers to anticipate possible problems and provide adaptations to the course or methodology, so instructors or even the institution can rethink the curriculum design or implement interventions to improve the learning experience.

It is very common for a post to be accompanied by a sequence of events, making it even more difficult to get the necessary attention from an instructor in a virtual learning environment. In these cases, students ask for colleagues' votes to put light on their posts and bring tutors' attention [Chaturvedi et al. 2014]. Therefore, the authors built models to predict the instructor's intervention using post votes, post-time variation, count of occurrences of words related to the evaluation, etc.

In [Marbouti et al. 2016], the instructors can use various strategies to communicate with students and lead them to paths that can improve their performance. For example, using a predictive system to identify students who are at risk and apply intervention guidelines so that the student can be successful. They highlight some shortcomings of academic early warning systems. One problem is that they typically employ a general prediction model that fails to address the complexity of all courses. Another problem is that most of these systems are designed for online courses, being too dependent on Course Management System (CMS) access data.

Some works make use of attributes that express student's sentiment to avoid the use of data that are directly linked to the environment. In [Yang et al. 2015], confusion is analyzed at different levels, having its impact verified on learning. The confusion experienced during the learning process is not always associated with negative outcomes. An immediate response clarifying the confusion can lead the student to overcome it, having a beneficial effect.

Sentiment analysis is very important in MOOC environments. According to [Moreno-Marcos et al. 2018a], identifying whether forum messages are positive or negative can give an insight into how students feel about the course, aiming to increase engagement and satisfaction.

Unlike [Moreno-Marcos et al. 2018a, Yang et al. 2015], we consider three attributes (sentiment, confusion, and urgency) to predict student's feelings. [Marbouti et al. 2016] suggest the use of predictive systems to identify students at risk and support them. In the works of [Chaturvedi et al. 2014, Moreno-Marcos et al. 2018b], the prediction is performed based on data produced by the platform, which ends up generating a platform-dependent solution. In our approach, we seek for a platform-independent solution, thus data produced by the platform are not used, only student messages. Based on the related work and previous research, we propose a system architecture to carry out the pedagogical intervention, aiming to help students and tutors by identifying subjective information present in student messages in virtual learning environments. We trained three automatic textual classification models to identify the students' educational moments based on their feelings. These models identify three subjective attributes: sentiment, confusion, and urgency.

3. Proposed Architecture

The so-called CAERS (Conversational Agent for Educational Recommendation System) architecture works to carry out pedagogical interventions in virtual learning environments. It consists of monitoring the forums to automatically identify subjective attributes that determine how the student is feeling, according to the attributes of sentiment, confusion, and urgency. These attributes are implicitly present in interactions through text messages. Pedagogical intervention represents a means of helping students not to create feelings of abandonment and demotivation, which can conduct them to drop out. The cases identified as the most critical demanding more attention are automatically alerted to tutors so that they can execute an individualized follow-up. In the most standard cases, the intervention is carried out automatically, through motivational, informative, or thank messages.

Our approach identifies the semantic patterns in student messages using ML-based classification techniques, considering earlier approaches [Capuano and Caballé 2019, Bóbó et al. 2019, Braz et al. 2019]. Sentiment, urgency, and confusion [Capuano and Caballé 2019, Capuano et al. 2021] are the attributes automatically identified by the machine learning step. Later, our solution stores the students' messages and their attributes in an ontology, making it possible to make inferences to detect the necessary pedagogical intervention [Rossi et al. 2021]. It is possible to define which action the agent will carry out based on these attributes, applying specific dialogue patterns. Figure 1 presents the layers and workflow of the proposed architecture to carry out the pedagogical intervention [Rossi et al. 2021].

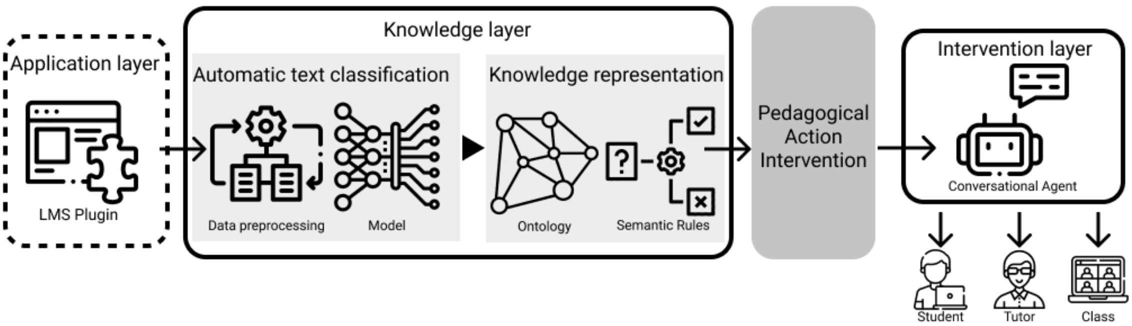

Figure 1. CAERS Architecture

The Application layer is responsible for capturing student interactions. This layer was designed to receive messages from learning management systems, such as posts on forums, and from the capture of these messages, it feeds the other layers. We have two modules in the knowledge layer: automatic text classification and knowledge representation. The textual knowledge classification module is of great importance, being responsible for identifying the subjective attributes present in the posts, such as Sentiment, Confusion, and Urgency. After classifying these attributes, they are directed to the knowledge representation module, where the student message data are stored, maintaining the same semantic knowledge. Finally, the necessary pedagogical intervention is identified based on the information stored, which can be an automatic motivation message, automatic help, tutor help, class help, or automatic thank message [Rossi et al. 2021].

The automatic text classification component is extremely important for the architecture, being responsible for labeling messages through predictive models. The models predict three attributes: sentiment, urgency, and confusion. The detection of these attributes does not represent a trivial task, as they are present in student messages, but in a subjective way, that is, none of the attributes is explicitly described. The attributes were selected because they better represent how the student feels, therefore contributing to the choice of the most appropriate pedagogical intervention.

These predictive models are built on training data enabling future predictions; that is, they analyze what pattern the data follow to serve as a predictive basis. The results achieved based on the models will serve to guide the choice of pedagogical intervention. The performance of predictive models is influenced by the quality of the data used in the training process. The preprocessing module is responsible for removing unwanted data, which does not add information for the models to the attributes prediction. Text preprocessing takes an input of raw text and returns cleaned tokens. Tokens are single words or groups of words [Anandarajan et al. 2019].

Once the data are preprocessed and organized, training the models starts using machine learning algorithms. The models are trained using the supervised training technique, where mapping is learned, given a set of input variables X and an output $\hat{Y}$ [Cunningham et al. 2008]. At the end of this training stage, we have three different models, one for each attribute: sentiment, confusion, and urgency.

Based on the trained models, it is possible to identify the subjective attributes present in the messages, that is, when submitting a message to the models, we obtain their respective classes. These attributes identified are loaded in the ontology and based on the semantic rules, the necessary pedagogical intervention is identified to support the student.

Taking into account the selected pedagogical intervention, the intervention layer can act in the student's learning process, intervening more effectively and appropriately. In this work, the intervention is carried out through dialog messages, but recommendations or diagnostic interventions can also be carried out.

4. Automatic Text Classification Evaluation

The proposed architecture was evaluated to measure the ability to automate students' message processing. This section evaluates the accuracy of the classification model for attributes based on messages that do not yet have a known label. One research question (RQ) was stated, and two secondary research questions (SRQ) were derived:

RQ1: How is it possible to automatically detect subjective information, which expresses how the student feels through his/her posts, using automatic textual classification techniques?

- SRQ1: Is it possible to detect subjective attributes independent of the platform information using only the post text?

This question intends to verify if we can obtain a classification model that does not depend on specific data coming from the platform, developing a solution that does not have the platform as a restriction.

- SRQ2: Do machine learning techniques detect sentiment, confusion, and urgency in students' post messages identifying how they are feeling?

This question aims to show whether the automatic text classification module proposed in the architecture can detect how students are feeling, classifying post messages based on sentiment, confusion, and urgency.

4.1. Materials and method

This work explored two different approaches to performing automatic text classification. The first used the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model and the second used Hereditary Tree-LSTM. A neural network performs attribute predictions from an input message in these approaches. Furthermore, BERT is pre-trained with unlabeled data for various tasks [Devlin et al. 2018] that can be fine-tuned to specific tasks.

The Tree-LSTM is an extension of the Long Short Term Memory (LSTM) standard, where the difference can be found in the organization of the LSTM unit cells, as they are structured in a tree format, passing information from the leaf nodes to the root node. The Tree-LSTM unit hierarchically incorporates information from each child node, while the standard LSTM disseminates historical information through a sequence of neural network units [Tai et al. 2015]. The Hereditary Tree-LSTM (HTL) contains a hereditary attention mechanism to help Tree-LSTM focus on relevant information about a natural language problem. This neural architecture achieves good results for semantic relationship tasks and text sentiment classification [Gomes et al. 2022].

For our work, the Stanford MOOCPosts[1] dataset was used. It is a dataset with a good amount of detail about the message and a reasonable volume of instances. Stanford MOOCPosts contains 29,604 anonymous student posts from eleven different content forums. The posts refer to courses in the areas of human sciences, medicine, and education. The messages in the dataset were labeled manually by three human coders [Capuano and Caballé 2019].

We compared our results to [Capuano and Caballé 2019, Capuano et al. 2021]. In their work, the authors discretized the attribute values into three classes. For the sentiment, values less than or equal to 3 correspond to the negative class. Values greater than 3 and less than 5.5 belong to the neutral class, and the positive class group values greater than 5. For the attributes of confusion and urgency, the first class has values less than 3.5. The second class has values greater than 3 and less than 5.5. Finally, the third class has values greater than 5. All messages passed through the preprocess step, being cleaned, removing texts between brackets, URLs, paths, and excess spaces to remove unnecessary information for the model training process and also reduce their dimensionality.

4.2. Results

For each instance, the models predict the class for each attribute, sentiment, confusion, and urgency. Precision, recall, and f-score metrics were analyzed to evaluate the models. The works [Capuano and Caballé 2019, Capuano et al. 2021] were used as a baseline, considering they also used the same dataset to train and evaluate our deep learning networks. The bow-ff approach was developed in [Capuano and Caballé 2019], and HAN was developed in [Capuano et al. 2021].

In [Capuano et al. 2021], the network with an attention mechanism obtained good results, indicating that using attention mechanisms can be an attractive option. BERT was chosen because it possesses several works in the literature that obtained good results, such

[1] https://datastage.stanford.edu/StanfordMoocPosts/

as [Khodeir 2021], which trained a model to recognize the urgency in MOOC course messages. Tree-LSTM was based on the approach presented in [Gomes et al. 2022], which also uses attention mechanisms.

Table 1. Experiment results

Attribute	Architecture	Epochs	Loss	Precision	Recall	F-Score
Sentiment	bow-ff	20	0.645	**88.62%**	86.07%	87.33%
	HAN	10	0.115	88.34%	88.29%	**88.31%**
	BERT	20	0.970	87.87%	**88.85%**	88.10%
	Tree-LSTM	20	1.477	84.70%	87.30%	84.80%
Confusion	bow-ff	20	0.081	87.19%	84.25%	85.70%
	HAN	10	0.103	87.25%	85.56%	86.40%
	BERT	20	0.352	**87.99%**	**86.09%**	**87.03%**
	Tree-LSTM	20	0.381	81.49%	85.76%	80.33%
Urgency	bow-ff	20	0.098	84.05%	75.75%	79.69%
	HAN	10	0.103	82.95%	76.40%	79.54%
	BERT	20	0.215	**86.84%**	**83.51%**	**85.12%**
	Tree-LSTM	20	0.455	77.20%	81.34%	77.98%

In Table 1, we can see that BERT and Tree-LSTM are numerically competitive with the baseline results. However, the results obtained by BERT showed slightly better results when compared to Tree-LSTM. When we look at the result achieved by BERT, we can see that sentiment is very close to HAN. When comparing confusion, BERT is slightly better than HAN; and when comparing urgency BERT is significantly better. The numbers show that it is viable to perform automatic text classification without using platform data. Therefore, based on Table 1, it is possible to answer SRQ1.

However, just looking at the numbers, it may not be the best analysis. Remembering that the objective is to help tutors and students, it is essential to correctly classify messages with a negative sentiment, as they identify the students who need more help and are certainly students who are unhappy with something. Therefore, it is very important to minimize errors in messages that are negatively biased.

The values of the sentiment attribute used to train the models, whose results were presented in Table 1, were discretized into three classes. However, we can increase the class numbers; we could have very negative, negative, neutral, positive, and very positive messages. According to the new discretization of values into five classes, we have the following distribution: sentiment values less than 2 represent a very negative polarization, values greater than 1.5 and less than 3.5 correspond to negative, greater than 3 and less than 5.5 represent neutral, those greater than 5 and less than 6.5 are positive, and those greater than 6 are very positive.

Table 2 compares the results obtained through discretization in 3 and 5 classes. Although the problem has become more complex, the metrics have not suffered great variations. However, the difference between the results obtained by BERT is smaller, while in Tree-LSTM, the difference between the results is a little bigger, so we can consider that BERT behaves better with the increase in the number of classes.

In order to evaluate the model obtained, we must explore the results better. In this

Table 2. Results using discretization of sentiment in 3 and 5 classes

Architecture	Epochs	Loss	Precision	Recall	F-Score
BERT (3 Class)	20	0.970	87.87%	88.85%	88.10%
BERT (5 Class)	20	1.070	86.29%	87.74%	86.84%
Tree-LSTM (3 Class)	20	1.477	84.70%	87.30%	84.80%
Tree-LSTM (5 Class)	20	0.598	75.61%	80.60%	76.66%

case, it is important to conduct qualitative analysis to observe the behavior of the classifier to know the real and predicted sentiment, especially of students who represent a negative sentiment, who are considered more critical to the problem of this work.

According to the result, the BERT model approach stood out from the others, so in-depth data analyses were carried out using the BERT approach. Checking the confusion matrix can help us evaluate more deeply and then know the model's behavior individually for each class. Even looking at what was misclassified also helps us better understand the model behavior. Figure 2 presents the BERT model behavior through the confusion matrix generated for the attribute sentiment divided into 5 classes.

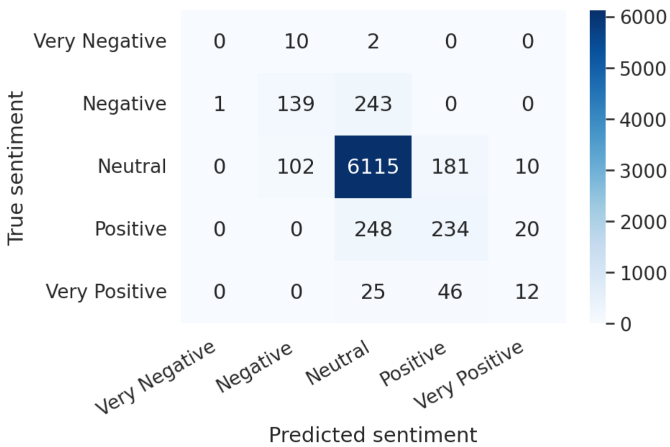

Figure 2. BERT model confusion matrix for 5 classes.

According to Figure 2, we can verify that with five classes, the sentiment classification was better compared to the classification for three classes. In addition, to recognize a greater number of messages with negative polarity, there was no classification of messages with negative polarity predicted to be positive. Therefore, as this work aims to help students and tutors, through pedagogical intervention, to avoid the feeling of abandonment and consequently combat evasion, we can consider that the classification with 5 classes was relatively better. Furthermore, as shown in Figure 2, we can assess that the model tends to classify messages as neutral. This is due to the imbalance of classes since the messages posted in the social interactions of virtual learning environments are normally not polarized about sentiment.

BERT has a limitation regarding the number of words, and we investigated whether this limitation is disturbing the model of correctly classifying messages. We check if there are considerable errors in messages that are above the word limit.

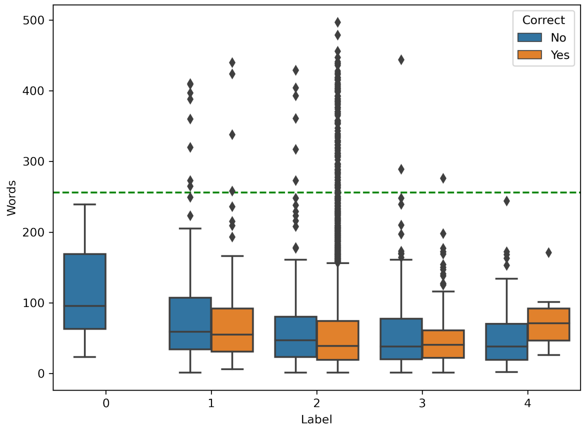

Figure 3. Chart presenting errors and successes distributed by class and post size

In Figure 3 we present a graph with the relationship between the size of the messages and the errors and successes. Boxes in blue represent incorrect classifications, while the correct classifications are in orange. The X-axis represents the different classes that represent sentiments, and the Y-axis determines the number of words in the message. Due to computational resource limitations, up to 100 tokens are considered per instance for training and evaluating the model. In the graph, the dotted line represents this limit in green color that separates the graph horizontally. As we can see, it is not possible to identify a relationship between incorrectly classified messages and their sizes. Therefore, limiting the maximum number of words allowed by the classifier cannot get in the way of the performance of the model.

Table 3. Messages obtained through the test set after sentiment classification. T stands for *target class* and P stands for *predicted class*. Very negative, negative, neutral, positive sentiments were labelled from 0 to 3, respectively.

Message	T	P
How sad is it that this statement is ever said. I hope to change this when I become a teacher.	1	0
When I was growing up, my third grade teacher always told our class that boys were better than girls in math. This perception had possibly impacted on how I performed as a student.	1	2
I was impressed by their comfort talking about math and familiarize with them	2	3
I was surprised by how the lads were so at ease with manipulating numbers to make them easier to work with. They enjoyed playing with numbers to get where they wanted as well, which is a concept I rarely see with students.	3	2
Students get negative messages about maths from teachers of other subjects. I have observed them rolling their eyes, shuddering, gleefully proclaiming how bad they are at math. A friend of mine observed an elementary school teacher threaten her students with math if they didn't behave.	2	3

For a better understanding of the errors made by automatic classification, Table 3 presents some samples in which instances were incorrectly classified. We can see that some messages leave doubts regarding the true sentiment because we do not know the criteria and methodology adopted to perform the manual annotation. In some cases, only after reviewing the annotation would it be possible to conclude that the model made an error in the classification, thus creating a strategy for the model to be more accurate.

Considering the analysis of the results obtained, we can say that it is possible to identify the students' feelings in the messages posted. Furthermore, using five classes, we were able to identify more messages with negative sentiment, characterizing a good alternative to identify students who need more immediate intervention. With this, we responded to SRQ2, making it possible to identify sentiment, confusion, and urgency in students' messages in virtual learning environments and perceive how they feel.

5. Final Remarks

The intervention of tutors and teachers is essential to support students in the teaching-learning process, answer questions about their content, and provide student engagement. In this work, we propose the CAERS architecture for automatic pedagogical intervention by identifying students' needs based on their messages in virtual learning environments.

This article focuses on developing the automatic textual classification module of the architecture. The messages are pre-processed and submitted to three models to identify sentiment, confusion, and urgency. These attributes define how the student feels and are useful to the choice of the most appropriate type of pedagogical intervention. The solution helps students and tutors, trying to make students feel more motivated and engaged.

In this work, some limitations were considered by us. About the annotation process, the evaluators' protocol is unknown to us, and there is a discrepancy between the values of some instances. So, the base may be subject to human classification errors. Therefore, there is no definitive research, and our results are not conclusive due to experiments have been carried out using only one dataset. However, we present the results to outperform the state of the art of the dataset. We know that the approach is domain-independent for testing other datasets, but it was merely for the English language. To test in other languages, it may be necessary to make some adaptations.

The evaluation was carried out using quantitative and qualitative analysis of the model results. This allows the general evaluation of the model and the individual evaluation of each class. It is worth noting that the correct classification of negative sentiment messages is very important for the proposed problem. The results show that the proposal surpasses other approaches found in the literature. In some cases, it was possible to obtain more significant results. In future works, we must revisit the process of manual message annotating to verify the errors made by the models compared with the criteria adopted during the annotation process. It is also important to carry out an assessment with stakeholders to analyze pedagogical interventions.

References

Anandarajan, M., Hill, C., and Nolan, T. (2019). Text preprocessing. In *Practical Text Analytics*, pages 45–59. Springer.

Bóbó, M., Campos, F., Stroele, V., David, J., and Braga, R. (2019). Identificação do perfil emocional do aluno através de análise de sentimento: Combatendo a evasão escolar. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1431.

Braz, F., Campos, F., Stroele, V., and Dantas, M. (2019). An early warning model for school dropout: A case study in e-learning class. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1441.

Capuano, N. and Caballé, S. (2015). Towards adaptive peer assessment for moocs. In *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, pages 64–69. IEEE.

Capuano, N. and Caballé, S. (2019). Multi-attribute categorization of mooc forum posts and applications to conversational agents. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 505–514. Springer.

Capuano, N., Caballé, S., Conesa, J., and Greco, A. (2021). Attention-based hierarchical recurrent neural networks for mooc forum posts analysis. *Journal of Ambient Intelligence and Humanized Computing*, 12(11):9977–9989.

Chaturvedi, S., Goldwasser, D., and Daumé III, H. (2014). Predicting instructor's intervention in mooc forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1501–1511.

Cuevas, R., Ntoumanis, N., Fernandez-Bustos, J. G., and Bartholomew, K. (2018). Does teacher evaluation based on student performance predict motivation, well-being, and ill-being? *Journal of school psychology*, 68:154–162.

Cunningham, P., Cord, M., and Delany, S. J. (2008). Supervised learning. In Cord, M. and Cunningham, P., editors, *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, pages 21–49, Berlin, Heidelberg. Springer Berlin Heidelberg.

Demetriadis, S., Caballé, S., Papadopoulos, P. M., Gómez-Sánchez, E., Kolling, A., Tegos, S., Tsiatsos, T., Psathas, G., Michos, K., Weinberger, A., et al. (2021). Conversational agents in moocs: reflections on first outcomes of the colmooc project. *Intelligent Systems and Learning Data Analytics in Online Education*, pages xxxvii–lxxiv.

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

Fandiño, F. G. E. and Velandia, A. J. S. (2020). How an online tutor motivates e-learning english. *Heliyon*, 6(8):e04630.

Gomes, J., de Mello, R. C., Ströele, V., and de Souza, J. F. (2022). A hereditary attentive template-based approach for complex knowledge base question answering systems. *Expert Systems with Applications*, page 117725.

Khanal, S. S., Prasad, P., Alsadoon, A., and Maag, A. (2020). A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25(4):2635–2664.

Khodeir, N. A. (2021). Bi-gru urgent classification for mooc discussion forums based on bert. *IEEE Access*, 9:58243–58255.

Marbouti, F., Diefes-Dux, H. A., and Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103:1–15.

Moreno-Guerrero, A.-J., Aznar-Díaz, I., Cáceres-Reche, P., and Alonso-García, S. (2020). E-learning in the teaching of mathematics: An educational experience in adult high school. *Mathematics*, 8(5):840.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., Estévez-Ayres, I., and Kloos, C. D. (2018a). Sentiment analysis in moocs: A case study. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1489–1496. IEEE.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., and Kloos, C. D. (2018b). Prediction in moocs: A review and future research directions. *IEEE Transactions on Learning Technologies*, 12(3):384–401.

Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., and Sindhi, S. (2018). Online education: Worldwide status, challenges, trends, and implications. *Journal of Global Information Technology Management*, 21(4):233–241.

Panigrahi, R., Srivastava, P. R., and Panigrahi, P. K. (2020). Effectiveness of e-learning: the mediating role of student engagement on perceived learning effectiveness. *Information Technology & People*, 34(7):1840–1862.

Rossi, D., Ströele, V., Campos, F., Braga, R., and David, J. M. N. (2021). Identifying pedagogical intervention in moocs learning processes: a conversational agent proposal. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 849–860. SBC.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Toti, D., Capuano, N., Campos, F., Dantas, M., Neves, F., and Caballé, S. (2020). Detection of student engagement in e-learning systems based on semantic analysis and machine learning. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 211–223. Springer.

Yang, D., Wen, M., Howley, I., Kraut, R., and Rose, C. (2015). Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the second (2015) ACM conference on learning@ scale*, pages 121–130.