## INTRODUCTION

Learning Management Systems (LMS) offer many features and tools that may be used as instruments of interaction between students and instructors. These instruments contribute to the learning experience because they support synchronous and asynchronous communication between all the participants in Distance Learning (DL) courses (Maciel, Rodrigues, & Carvalho, 2014). One of the most used features in these environments is the discussion forum. According to the Census EaD.br (2016), the discussion forum is a tool widely adopted by educational institutions as a way to communicate with students, being used approximately 72% of Brazilian distance learning courses.

In discussion forums, the number of posts at the end of a debate varies according to the theme, the interest of participation provoked by the topic, and even the profile of the students participating. Parts of the postings may contain interesting arguments or positions on the topic of discussion, such as citation to important concepts or arguments linking different concepts. On the other hand, there are often posts that do not contribute significantly to the discussion, including information that is not relevant to the topic.

The involvement of students in the forums is an important activity for the construction of knowledge and generally a sign of students interest. Through the analysis of students' interactions in the tool, instructors can identify learning difficulties. Yet, due to the large number of messages that are created, a manual analysis of student posts can become a very time-consuming task. Furthermore, the analysis of the written content in forums may free instructors to spend more time focusing on causes that may be leading to possible lack of participation by students (Azevedo, Behar, & Reategui, 2011).

In this context, approaches that use Text Mining (TM) are becoming very promising in the process of knowledge discovery, by using techniques of analysis and extraction of data from unstructured texts (Morais & Ambrósio, 2007). More specifically, an option that may be considered together with Natural Language Processing (NLP) is Network Science (NS) (Albert & Barabási, 2002; Newman, 2003), a concept that comes from statistical mechanics and makes extensive use of graph theory algorithms. The appropriate treatment of texts can provide relevant information about students, as well as their needs, knowledge, and interests (Panceri & Menezes, 2007).

Therefore, the objective of this work is to introduce an approach for modeling and development of a solution that allows the thematic relevance analysis of posts written by students in discussion forums. This approach prioritizes the identification of posts of students who may be facing difficulty in learning, while making sure the relevant contributions on the topic of the discussion are also identified. We developed a computational model that seeks to evaluate the thematic relevance of the posts made by students in relation to the opening text of the forum, which we assume to define the problematization of the discussion.

Besides this introduction, this article is organized into five other sections which present the related works, the proposed workflow, the experiments, the results and discussions, and the final considerations.

## RELATED WORKS

Several scientific studies have been carried out in DL environments with a focus on the investigation of aspects associated with both the communication and the interaction of students in LMS (Chen & Looi, 2017; Machado, Lima, Maciel, & Rodrigues, 2016). The activities related to

the communication and interaction of students in these virtual environments promote the Computer-Supported Collaborative Learning (CSCL) which, according to Bogarín, Cerezo, and Romero (2018), *"is characterized by the sharing and construction of knowledge between participants using technology as their primary means of communication or as a common resource"* (p. 10). From the computational point of view, several techniques have achieved promising results, among them Text Mining and Network Science deserve our special attention as they are related to the scope of this work.

## Text Mining

According to Tan (1999), Text Mining (TM) performs the process of extracting non-trivial patterns from texts. Inspired by data mining, which seeks to discover emerging patterns in structured data, TM aims to extract useful insights from unstructured or semi-structured data (Aranha & Passos, 2006). This process, however, can be accomplished through a variety of techniques, as in Li and Huang (2008), where a multidimensional analysis approach was proposed to investigate the contributions of students through a model that represents the texts in a vector space. In this technique, each document term becomes a dimensional characteristic. It can be seen, however, that the vector representation discards important information, such as the order in which the terms appear, where they appear, and the proximity between them (Schenker, 2003).

Another approach considers representing texts through graphs to overcome the aforementioned limitation. In this approach, the nodes of the graph generally represent the most frequent words whilst the associations between the nodes of the graph indicate the proximity between the words. In the context of graph representations we can highlight three interesting works. N. Chen, Kinshuk, and H. Chen (2008) developed metrics to quantify the strength of the relationship between keywords identified in a set of texts; Lee and Segev (2012) proposed a mechanism for the automatic generation of conceptual maps in e-learning from textual documents; and Azevedo (2011) used Text Mining through graphs to develop a resource for qualitative analysis of textual contributions recorded by students in discussion forums.

## Networks Science

Network Science (NS) is a new scientific field based on a network representation of of real-world phenomena where the network tends to present non-trivial properties (Barabási, 2003). These networks have properties that can be analyzed as structural measures and are often used to better understand the phenomena they represent. Amongst the metrics that are potentially useful in the extraction of characteristics in textual contents, we can highlight the betweenness centrality (Freeman, 1977).

Betweenness is a measure of centrality which considers that a vertex is important if it is part of a large number of minimum paths. High values of betweenness indicate that the vertex has high influence in the network by the mere fact that information flowing in the network tend to pass through these nodes (Medeiros, 2015). In the context of textual analysis, words whose nodes in the network have high values of betweenness are those that have high frequency and also those that connect other words or concepts, being able to indicate aspects about the context of the written content.

Numerous works in the literature use NS to extract information into written contents. Antiqueira, Nunes, Oliveira, and Costa (2005), for example, have modeled texts as complex networks to perform an evaluation of the quality of written content. In their model, a text is represented by a network in which each word is a vertex and each edge represents co-occurrence

of the words. In the work performed by Amancio (2013) NS was used to treat problems of textual classification in order to identify patterns that relate the structure of the text verified by the metrics of networks with their style's characteristics of the written text. Al Rozz and Menezes (2018) have applied some approaches based on NS to extract textual features through topological properties with the intention of identifying the authorship of the written contents.

## PROPOSED ARCHITECTURE

This section presents the architecture of our proposed Text Processing Engine. For this research, a database with real-world records of a LMS was used. The texts extracted from the LMS forums were written by students and instructors in the Brazilian Portuguese language. Figure 1 presents the general architecture of the processing engine; we have used Python during the development process.

**Figure 1. Text Processing Engine - General Architecture.**

The Processing Engine is initialized with the identifier of the forum to be analyzed. With this parameter, it is possible to extract all the contents of the forum in the database through SQL queries. Then, for each post, the following modules are executed sequentially: Preprocessing; Keywords Extraction; creation of the Knowledge Map; and the evaluation of Thematic Relevance. These modules perform specific tasks in the process of extracting knowledge and are described in detail below. At the end of the execution, a table containing the data of all processed posts is outputted.

### Preprocessing

The preprocessing module is responsible for handling all textual content that has been stored in the LMS database. This module receives the data from a post extracted from the database as input. Figure 2 illustrates the steps performed during text preprocessing.

**Figure 2. Steps of the preprocessing module.**

The proposed approach starts by looking if the users' post is coherent with the theme of the original post inserted by the instructor at the time of opening the forum. This information is important in capturing the content of the discussion since the instructor can add in the post instructions and general rules of participation in the forum, which should not be considered as part of the discussion context. If the post is outside the theme set by the instructor, the topic of the forum is extracted; otherwise, the textual content of the post is captured in full.

In the next step, any links in the text are removed. In general, links to Web pages are used as a way of referring to the content of an argument or, as in some cases, include suggestions for further reading. After this, the content of the post is passed through a process of *lemmatization* that is responsible for transforming the words into their respective canonical forms, enabling the representation of each word of the text in a unique format, regardless of gender, tense, or number variations. In this research, the lemmatization was performed through the Cogroo API, which is an open-source grammar checker for the Brazilian Portuguese language. This tool performs the hybrid analysis considering statistical techniques of natural language processing and a rule-based approach (Silva, 2013).

After the lemmatization, the sentences are separated from the text. In the next step, the words in each sentence are tagged morphosyntactically through the Cogroo API. Finally, the stopwords

are removed considerering the morphosyntactic class assigned to each word. In this step, we argue that only words classified as verb, noun, adjective or adverb should be considered. All other words, such as articles, prepositions, numerals, among others, are discarded.

## Keywords Extraction

The module responsible for performing the keywords extraction receives the pre-processed text as input and then executes a sequence of activities in order to identify the main terms for the context of the argumentation. Figure 3 shows the steps performed during the extraction of the keywords.

**Figure 3. Steps of the keyword extraction module.**

The first step is in charge of indexing the words in the text. This step is characterized by uniquely identifying each word in the text, as a way of generating a list where each position corresponds to a different word. In the next step, an adjacency matrix containing the co-occurrence values between the words in the text is generated; the co-occurrence value refers to the number of times two words appear together in the text.

Once the adjacency matrix has been generated, it represents the network of words for that particular text. In this network, the nodes represent the words, while the edges between the nodes indicate that there is an adjacency relationship between the word pairs. We then calculate the value of the betweenness centrality for each node in the network. This metric assumes that a vertex is important if it part of a large number of minimum paths.

The betweenness value of a node $v$ is calculated as the sum of the fraction of all pairs of minimum paths passing through $v$, represented by Equation (1):

$$C_B(v) = \sum_{s,t \in \mathcal{V}} \frac{\sigma(s,t)}{\sigma(s,t|v)} \quad (1)$$

where $\mathcal{V}$ is the set of network nodes, $\sigma(s,t)$ is the number of minimum paths between $s$ and $t$, and $\sigma(s,t|v)$ is the number of minimum paths between $s$ and $t$ passing through $v$.

The value of the betweenness is used to quantitatively indicate the important nodes of the network. Based on the context of the content written in the post, through this measure, it is possible to identify the words with greater importance in the textual structure. In this way, each position of the list with the keywords stores the information in the format "(word, morphosyntactic class, betweenness)", this list is organized in descending order by the value of the betweenness; the betweenness of each node is normalized from 0 to 1. The closer to 1 the value of the betweeness of a word, the more important it is in the text. Conversely, when approaching 0, it is assumed that the word becomes less relevant to the center of the topic in the forum.

## Knowledge Map

After identifying the keywords, the knowledge map module receives the tagged sentences which represents the preprocessed text, and the list containing the keywords. Knowledge maps are important learning resources because they enable the user to recognize important concepts and their relationships (Lee & Segev, 2012). These maps have nodes as important concepts and edges to associate the concepts that are related. For pairs of keywords, relationships are established considering the number of appearances in a sentence and the total number of words in the sentence. In this way it is possible to store the information that identifies the force of relation between the keywords as edges in the graph.

We take into consideration that the strength of the relationship between pairs of keywords is inversely proportional to the size of sentences, based on Lee and Segev (2012). In this way, as the number of words in a sentence increases, the weight of a word in the sentence decreases. This means that the relationship between two keywords in a shorter sentence is stronger than in longer sentences. Thus, to calculate the force of relation between two keywords in a text, we use the Equation (2):

$$R(k_i, k_j) = \sum_{t=1}^{n} \frac{2}{len(S_t)} \quad (2)$$

where $R(k_i, k_j)$ is the value of the relationship between the pair of keywords $k_i$ and $k_j$, $n$ are the sentences that contain these keywords and $len(S_t)$ is the length of the sentence $S_t$. The numerator "2" in this Equation represents the fact that in the sentence $S_t$ there are at least 2 keywords: the pair of keywords $k_i$ and $k_j$.

After calculating the relationship strength for all keyword pairs, the network representing the text knowledge map is created. In this network, the weight of the edges directly reflects the relationship between the words and is represented in the picture with thicker edges. Similarly, the node size, which is obtained from betweenness, characterizes the importance of the word to the context of the written content. Figure 4 exemplifies the process applied in the text of a student's post to generate the map of knowledge.

**Figure 4. (a) Original Text; (b) Pre-processed Text; (c) Resulted Network for Knowledge Map.**

## Thematic Relevance

Once the knowledge map creation is complete, the thematic relevance analysis module is executed. This module is inputed with the networks that represent the knowledge maps, both for the text of the post being processed and for the text of problematization. The purpose of this step is to identify if the text in question is related to the context of the topic proposed at the opening of the forum.

The thematic relevance analysis is done based the work of Azevedo et al. (2011), where a qualitative analysis in discussion forums was proposed, using text mining and comparison of the graphs formed from the contributions of the students. The authors defined the Thematic Relevance Coefficient (TRC) of a textual contribution as a measure able to represent how much a text is relevant in relation to the topic of discussion.

The TRC indicates the degree of relevance of a post according to the topic of the forum, which is calculated by the number of relevant concepts used in the post text that corresponds to relevant concepts in the problematization text, added to the number of associations correctly made between these relevant concepts. Thus, to measure the thematic relevance of a textual contribution, we analyze the equivalences between the nodes and the edges of the formed networks.

As argued by Azevedo et al. (2011), we assume that a text is relevant to the topic of the discussion if it presents concepts and associations equivalent to the topic of the debate. However, we believe that analyzing in isolation the number of nodes and edges that are common to both the networks generated by the student's text and the one generated by the problematization text, can omit important information regarding the degree of importance of each word or association to the context of the debate. This approach disregards the fact that words and their associations in a discourse have different weights (Lee & Segev, 2012).

In this way, when analyzing the network generated by the student's post, we consider the weight of each node and edge. The weight of these elements (nodes and edges) represents their importance in the network. When identifying an element in the complex network of the student's text ($N_e$) that has an equivalent in the network of the problematization text ($N_p$), the coefficient of equivalence (CE) is calculated according to Equation (3):

$$\mathrm{CE}(e, \wp) = W_{(\wp, N_p)} \times 2^{-|W_e - W_\wp|} \tag{3}$$

where $\mathrm{CE}(e, p)$ refers to the coefficient of equivalence between the elements $e$ and $\wp$, where $e \in N_e$ and $\wp \in N_p$. $W_e$ and $W_\wp$ represent the weight of the elements $e$ and $\wp$, respectively. The weight of all elements in each of the networks is normalized in a scale from 0 to 1. $|W_e - W_\wp|$ makes it possible for CE to vary according to the similarity between the weight of the elements $W_e$ and $W_\wp$. Thus, as the value of $W_e$ gets close $W_\wp$, the term $|W_e - W_\wp|$ tends to zero. On the other hand, the larger the difference between the weights of the elements $W_e$ and $W_\wp$, the closer to 1 is the value of $|W_e - W_\wp|$, which causes the power result to approximate 0. $W_{(\wp, N_p)}$ denotes the weighted weight of the element $\wp$ in relation to all elements of the same nature, i.e. nodes or edges, in $N_p$, which is calculated by Equation (4):

$$W_{(\wp, N_p)} = \frac{W_\wp}{\sum_{i \in N_p} W_i} \tag{4}$$

This approach to the calculation of CE allows for the weight of the element to influence the result of the equivalence. Thus, nodes and edges with higher weights tend to generate larger coefficients, while those with smaller weights are more likely to exhibit the inverse behavior. In addition, another factor incorporated is the similarity between the weights of the pair of nodes or equivalent edges. This feature makes it possible to argue that the closer the value of the weights of the pair analyzed, the more similar they will be, which means that the influence in the context is also similar. Algorithm 1 presents the pseudo-code of the thematic relevance analysis of a post.

**Algorithm 1. Algorithm for the thematic relevance analysis.**

```
Inicialize CE-Nodes and CE-Edges with 0;
sumNodes = Sum weights' nodes (N_p);
sumEdges = Sum weights' edges (N_p);
for each nodeE in getNodes(N_e) do
  for each nodeP in getNodes(N_p) do
    if isEquivalent(nodeE, nodeP) then
      Calculate CE(nodeE, nodeP) according to Equation 3;
      Increase CE-Nodes;
      for each nodeNeightorE in getNeighbor(nodeE) do
        for each nodeNeightorP in getNeighbor(nodeP) do
          if isEquivalent(nodeNeighborE, nodeNeighborP) then
            edgeE = getEdge(nodeE, nodeNeighborE);
            edgeP = getEdge(nodeP, nodeNeighborP);
            Calculate CE(edgeE, edgeP) according to Equation 3;
            Increase CE-Edges;
return (CE-Edges, CE-Nodes);
```

The equivalence check considers that two nodes are equivalent if they are synonymous to each other and belong to the same morphosyntactic class. The use of synonyms is an important feature to be considered when comparing the networks. If the student's post's network has any node that is synonymous with a node in the problem text, they will be considered equivalent. In this study, the TeP 2.0 (Maziero, Pardo, Di Felippo, & Dias-da-Silva, 2008) – an Electronic Thesaurus for Brazilian Portuguese, was used to do the synonymy analysis. This thesaurus is an electronic dictionary of synonyms and antonyms for the Brazilian Portuguese language.

After the thematic relevance analysis, the calculated values for CE-Edges and CE-Nodes are inserted into an external table, together with the post identifier. This table, which represents the knowledge base obtained as output from the text processing engine, is responsible for storing the data for all posts in the same forum.

## EXPERIMENTS

To evaluate the methodology proposed in this work, we carried out experiments in a database with real records of the learning management system used by the nucleus <*omitted review*> at the university <*omitted review*>. This section describes these experiments based on the organization of the scenarios that have been explored. We then report on the transformations and statistical tests that were applied to the resulting Textual Processing Engine data. Finally, we present the classification algorithms and the evaluation metrics we used in the processed data.

### Scenarios Organization

The experiments were performed in four forums of a graduate course in Pedagogy due to the high level of student participation in the forums, which in turn leads to discussions with a larger number of posts. Table 1 summarizes the general characteristics of the forums used in the experiments.

**Table 1. Forums characteristics. Note that the themes listed are in Brazilian Portuguese and are presented just for informational purposes. The specific translation of each theme is not relevant for this paper.**

| Course | Period | Forum Theme | Total Posts |
|---|---|---|---|
| Developmental Psychology | 1st | "*1º Fórum Temático: 'Pau que nasce torto morre torto'*" | 170 |
| | | "*3º Fórum Temático: 'Não vou me adaptar'*" | 171 |
| Art Education | 5th | "*1º Fórum Temático: A história do ensino de arte no Brasil*" | 197 |
| | | "*3º Fórum Temático: Metodologia do ensino de arte*" | 210 |

Each post made by the students in the forums are graded (evaluated). These grades are assigned by the instructors and rank the students' posts by a score ranging from 1 (minimum grade) to 5 (maximum grade). In this way, we use SQL scripts to extract the evaluation grade given by the instructor to each post and we added this information to the table along with the processed data of the text. The inclusion of the instructors' scores allowed us to run experiments in two specific scenarios intended to validate the methodology proposed for the classification of the postings regarding the thematic relevance.

In the first scenario, named 4F748P (which represents a dataset that consists of 4 forums with a total of 748 posts), we considered all the posts evaluated (the sum of collum "Total Posts" in Table 1). In this scenario, for the posts with an instructor score equal to 1, 2, or 3, we added the label "little relevant". On the other hand, the posts that received scores equal to 4 or 5 were labeled "relevant".

In the second scenario, named 4F313P (4 forums with a total of 313 posts), we considered only the posts with the worst evaluations, i.e. with the grade equal to 1 or 2, and the best-evaluated

posts, i.e. with the grade equal to 5, which represent the posts that received maximum marks from the instructor. In this scenario, the texts with the lowest scores were labeled “little relevant”, while those with a maximum score were labeled “relevant”.

## Transformations and Statistical Tests

After the definition of the scenarios and the transformation of the grade variable, we applied a linear normalization process in the attributes CE-Edges and CE-Nodes. This transformation carried out the scaling of these attributes for each of the forums separately, in order to avoid the influence of the records with values of greater magnitude. Thus, normalized data correspond to the interval [0,1].

In the next step, the correlation between the CE-Edges and CE-Nodes metrics was calculated in relation to the Instructor’s Grade (TG). For this purpose, the polyserial correlation (numeric-ordinal) (Siegel & Castellan, 1975) was applied according to the nature of the variables, where CE-Edges and CE-Nodes are continuous variables and TG is a scalar variable.

For the first scenario (4F748P), the correlations were (CE-Edges, TG = Cor. 0.4688) and (CE-Nodes, TG = Cor. 0.5390). For the second scenario (4F313P), the correlations were (CE-Edges, TG = Cor. 0.4234) and (CE-Nodes, TG = Cor. 0.4990). In Figure 5 the CE-Nodes metric boxplot for both scenarios is presented in relation to the TG, which is represented by the performance of the post.

**Figure 5. Boxplot of CE-Nodes as a function of TG for both scenarios.**

The calculation of the correlations, as well as the boxplot graphs, are only descriptive measures. Thus, it was necessary to apply a statistical test in order to verify the significance of the relationship between CE-Nodes and TG. For this, we used the non-parametric Kendal test (Siegel & Castellan, 1975). For the first scenario (4F748P), when we used the CE-Nodes and TG metrics we obtained a $p$-value = 0.04022, and for scenario 4F313P we obtained a $p$-value = 0.02835. This shows that, for both scenarios, there are significant correlations between the CE-Nodes metric and the instructor-assigned grade for the student's post.

This statistical analysis was important to encourage the use of classification algorithms capable of predicting student's grade based on CE-Edges and CE-Nodes values.

## Classification Algorithms

In this step, we used five supervised algorithms, to perform the classification of the posts in relation to the thematic relevance. The algorithms were: Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and K-Nearest Neighbors (KNN). All the algorithms tested in this work were used through the standard configuration provided by the Weka 3.8 tool.

As there are two previously defined classes, which are “little relevant” and “relevant”, we have a binary classification problem, with the number of samples equal to 748 and 313, represented by the total of posts in the 4F748P and 4F313P scenarios, respectively. The set of characteristics used by the classifiers is composed of the attributes CE-Edges and CE-Nodes, which represents a two-dimensional space.

For the MLP, the number of hidden layers was determined by the average between the number of attributes (CE-Edges and CE-Nodes) and the number of classes ("little relevant" and "relevant"), which totaled two layers. The learning rate set in this algorithm was 0.3. For the SVM, the parameter C was established in 1.0 and the kernel used was the polynomial function.

For the KNN, the value of k = 1 was adopted, which means that the classification will consider only one neighboring register. The DT algorithm was configured with a confidence fact of 0.25 and the minimum number of instances per sheet was equal to 2. For RF, the value 0 was used for the maxDepth parameter, which indicates that the depth of the tree is not limited.

In the scenarios, we adopted the k-fold cross-validation method, with the value of k equal to 10, which means that the data were divided into 10 subsets, of which 9 are used for the classifier training, while the subset was used for testing. The permutation of all subsets composes a simulation. For each algorithm used, 100 simulations were performed, and the result considered was calculated by the average value of all the values obtained in each scenario evaluated.

## RESULTS AND DISCUSSION

To evaluate the performance of the classifiers when processing the data extracted from the proposed model, we used the following metrics: Accuracy, Precision, Recall and *f*-Measure. We adopted the posts labeled as "little relevant" as the main class, in order to investigate the behavior of algorithms when dealing with posts that may possibly be associated with students with learning difficulties. Although posts classified as "relevant" also present information important to the instructor in decision making, identifying posts with low relevance contributes significantly to the teaching-learning process, due to the difficulty faced by instructors in perceiving students in need of pedagogical support at early stages of delivering the instruction.

### Evaluation of Classification Models

The results shown in Figure 6 represent the values of the performance metrics that were extracted from the classifiers in the 4F748P case. In this scenario, the "little relevant" class refers to posts with grades equal to 1, 2 or 3 (according to the instructor's assessment) which totals 392 texts. The "relevant" class contains the posts evaluated by the instructor with grade 4 or 5, which represents a total of 356 texts.

**Figure 6. Classification Performance - Scenario 4F748P.**

When we observed the result of the classification in relation to *accuracy*, we noticed that the best results were obtained by the algorithms MLP, DT, and SVM, with an accuracy rate of approximately 73%. This result points to a satisfactory general classification index since this metric considers the labeling obtained experimentally for both the "not relevant" class and the "relevant" class.

When evaluating the values obtained in relation to *precision*, we noticed that the MLP and DT algorithms excelled with a hit rate of 79%. With this, we can observe that there was a better performance of these classifiers when considering the positive predictive value, which is characterized by the correctness index of the posts that were correctly classified as "little relevant". On the other hand, when analyzing the behavior of the algorithms when dealing only with posts of the "little relevant" class, which is represented by *recall* metric, the best result was reached by the algorithm SVM, which obtained a 75.25% hit rate. Likewise, when we consider the incorrectly classified posts for both "little relevant" and "relevant" classes, which is obtained by the *f-measure* metric, the best performance remained associated with the SVM, with 73.45% of texts correctly classified.

For the scenario 4F313P, we disregard the posts with intermediate notes, which are those that received grade 3 or 4 according to the instructor's evaluation. In this configuration, we seek to investigate the performance of the algorithms in dealing with only the most heterogeneous texts

on the aspect of thematic relevance. For this, the class “little relevant” contemplates the posts with grades 1 or 2, which is composed of 140 texts. The “relevant” class consists of the posts evaluated by the instructor with grade 5, which contains 173 texts. Figure 7 shows the performance obtained by the classifiers for the metrics used.

**Figure 7. Classification Performance - Scenario 4F313P.**

As in the previous scenario, we noticed that the best results when considering the hit rate for both “little relevant” and “relevant” class are associated with the MLP and SVM algorithms, which reached approximately 84% *accuracy*. These algorithms stand out due to the correct classification of the posts as "non-relevant", represented by *precision*, with emphasis on the 86.2% success rate obtained by MLP.

On the other hand, when considering only the correct classification of the posts contained in the class “little relevant”, which is obtained by *recall*, the SVM performed the best (80%). It is also worth pointing the results achieved by this algorithm when considering the posts incorrectly sorted for both classifications; the success rate was 81.87%, which is reflected by the *f-measure*.

## FINAL CONSIDERATIONS

This paper proposes an approach to analyze the thematic relevance of the postings made by students in the discussion forums in distance learning environments. This proposal sought to develop a pedagogical support resource for the instructor, in the sense of early identification of students with learning difficulties, as well as students who have made relevant contributions on the topic of the discussion. To achieve these objectives, text mining and networks science techniques were applied to the processing of the posts. Then, the processed data was evaluated in five supervised learning classification algorithms.

We conclude that the approach used in this work allowed us to obtain promising results in the identification of the posts regarding their thematic relevance. These results point to a good level of success of the proposed methodology, which was reflected in a considerable performance of the classifiers. The results obtained by the proposed model are appropriate to be applied in real environments, given that they can provide the instructor with important information about students learning process, helping the instructors make better decisions.

As future works, we plan to incorporate features that consider the similarity between students' posts, in order to perceive the cases in which the text analyzed is analogous to some previous post. In addition, we intend to include in the processing engine the perception of the evolution of the context of the discussion, as new relevant posts are inserted. These initiatives seek to improve the performance of textual processing, which is reflected in a higher rate of accuracy in the classification of posts.

## ACKNOWLEDGEMENTS

This work was supported by the nucleus <*omitted review*>, department of distance learning at university <*omitted review*>.

## REFERENCES

Al Rozz, Y., & Menezes, R. (2018). Author Attribution Using Network Motifs. In: Cornelius, S.; Coronges, K.; Gonçalves, B.; Sinatra, R.; Vespignani, A. (eds) Complex Networks IX. CompleNet 2018 (pp. 199-207). Springer Proceedings in Complexity: Springer, Cham.

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. Rev. Modern Phys., 47-97.

Amancio, D. R. (2013). Classificação de textos com redes complexas. São Carlos, Brasil: Instituto de Física de São Carlos - USP.

Antiqueira, L., Nunes, M. G., Oliveira, O. N., & Costa, L. F. (2005). Modelando textos como redes complexas. Anais do III Workshop em Tecnologia da Informação e da Linguagem, (pp. 22-26). São Leopoldo - RS, Brasil.

Aranha, C., & Passos, E. P. (2006). A tecnologia de mineração de textos. RESI – Revista Eletrônica de Sistemas de Informação, 66-73.

Associação Brasileira de Educação a Distância - ABED. (2016). Censo EaD.br: relatório analítico da aprendizagem a distância no Brasil 2015. Curitiba - Brasil: InterSaberes.

Azevedo, B. F. (2011). MineraFórum: um recurso de apoio para análise qualitativa em fóruns de discussão. Porto Alegre, Brasil: Universidade Federal do Rio Grande do Sul - UFRGS.

Azevedo, B. F., Behar, P., & Reategui, E. (2011). Qualitative analysis of discussion forums. International Journal of Computer Information Systems and Industrial Management, 671-678.

Barabási, A. L. (2003). Linked: How everything is connected to everything else and what it means for business, science and everyday life. New York, United States: Basic Books (AZ).

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

Chen, N. S., Kinshuk, C. W., & Chen, H. J. (2008). Mining e-learning domain concept map from academic articles. Computers & Education, 1009-1021.

Chen, W., & Looi, C. K. (2017). Measuring process and outcome of the scientific argumentation in a CSCL environment. W. Chen, J.-C. Yang, A. F. Mohd Ayub, S. L. Wong, & A. Mitrovic (Eds.), Proceedings of the 25th International Conference on Computers in Education, 139-145.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry, American Sociological Association, 35-41.

Lee, J. H., & Segev, A. (2012). Knowledge maps for e-learning. Computers & Education, 353-364.

Li, Y., & Huang, R. (2008). Analyzing peer interactions in computer-supported collaborative learning: model, method and tool. LNCS - Lecture Notes in Computer Science, 125-136.

Machado, C. J., Lima, B. R., Maciel, A. M., & Rodrigues, R. L. (2016). An investigation of students' behavior in discussion forums using educational data mining. The 28th International Conference on Software Engineering & Knowledge Engineering - SEKE, (pp. 510-514). California, United States.

Maciel, A. M., Rodrigues, R. L., & Carvalho, E. C. (2014). Desenvolvimento de um assistente virtual integrado ao Moodle para suporte a aprendizagem online. Anais do XXV Simpósio Brasileiro de Informática na Educação, (pp. 382-391). Brasil.

Maziero, E. G., Pardo, T. A., Di Felippo, A., & Dias-da-Silva, B. C. (2008). A base de dados lexical e a interface web do TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil. Workshop em Tecnologia da Informação e da Linguagem Humana (pp. 390-392). Vila Velha, Brasil: XIV Simpósio Brasileiro de Sistemas Multimídia e Web.

Medeiros, M. A. (2015). Reconhecimento de autoria de textos utilizando redes complexas. Recife, Brasil: Universidade de Pernambuco – UPE.

Morais, E. A., & Ambrósio, A. P. (2007). Mineração de textos. Goiás - Brasil: Instituto de Informática da Universidade Federal de Goiás.

Newman, M. E. (2003). The structure and function of complex networks. SIAM Review, 167-256.

Panceri, S., & Menezes, C. (2007). Apoio computacional para a mediação pedagógica em 'Debate de Teses'. Goiás - Brasil: Instituto de Informática - Universidade Federal de Goiás.

Schenker, A. (2003). Graph-theoretic techniques for web content mining. Florida, United States: University of South Florida.

Siegel, S., & Castellan, N. J. (1975). Estatística não-paramétrica para ciências do comportamento. São Paulo, Brasil: Artmet Editora.

Silva, W. D. (2013). Aprimorando o corretor gramatical CoGrOO. São Paulo, Brasil: Instituto de Matemática e Estatística, Universidade de São Paulo.

Tan, A. (1999). Text mining: the state of the art and the challenges. Workshop on Knowledge Discovery from Advanced Databases, (pp. 71-76). Beijing.