Understanding CSCL Peer Feedback Contributions: An Automated Content Analysis Approach

Mayara Simões de Oliveira Castro
Federal Rural University of Pernambuco
Recife, PE, Brazil
maaycastro@hotmail.com

Rafael Ferreira Mello
Federal Rural University of Pernambuco
Recife, PE, Brazil
rafael.mello@ufrpe.br

ABSTRACT

Peer feedback has been widely used in computer-supported collaborative learning (CSCL) setting to improve students' engagement with massive courses. Although the peer feedback process increases students' self-regulatory practice, metacognition, and academic achievement, instructors need to go through large amounts of feedback text data which is much more time-consuming. To address this challenge, the present study proposes an automated content analysis approach to identify relevant categories in peer feedback based on traditional and sequence-based classifiers using TF-IDF and content-independent features. We use a data set from an extensive course (N = 231 students) in the setting of engineering higher education. In particular, a total of 2,444 peer feedback messages were analyzed. The results have shown promising outcomes with both TF-IDF and content-independent features. The Conditional Random Fields (CRF) classification model based on the TF-IDF features achieved the best performance, considering all the metrics computed in the analysis. The results illustrate that the ability to scale up the automatic analysis of peer feedback provides new opportunities for student improved learning and improved teacher support in higher education at scale.

CCS CONCEPTS

• **Applied computing** → **Collaborative learning**; • **Information systems** → **Content analysis and feature selection**.

KEYWORDS

Content Analysis, Peer Feedback, Natural Language Processing, Computer Supported Collaborative Learning.

ACM Reference Format:
Mayara Simões de Oliveira Castro and Rafael Ferreira Mello. 2023. Understanding CSCL Peer Feedback Contributions: An Automated Content Analysis Approach. In *Proceedings of Annual ACM Conference on Learning at Scale (L@S2023).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

L@S2023, July20–22, 2023, Copenhagen, Denmark

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Collaborative learning has significant importance in the education system. It offers students opportunities to learn valuable interpersonal and teamwork skills by participating in task-oriented learning groups [7]. Research has demonstrated that collaborative learning influences student academic achievement (e.g., [48, 56, 58]). However, we still have a limited understanding of the nature of student regulation of learning in computer-supported (CSCL) settings, especially in large courses.

In this context, peer assessment and peer feedback are key components in the collaborative learning setting [27]. Peer assessment was shown to increase learning performance [32]. Peer feedback, in particular, can also be an effective way to help students reflect on their contributions and identify areas for improvement [34]. However, it is time-consuming and often impossible to keep track of the feedback given by each student in a group in a large-enrollment course (more than 100 students) [54]. Therefore, there is a need to develop tools and methods that would allow for automated feedback analysis to improve teacher support and, ultimately, learning.

Recent studies have applied machine learning methods to automate content analysis in educational settings [12, 19, 20]. More specifically, the results of the automated content analysis can be used to analyze student/instructor feedback to support better learning practices [11, 12]. Nevertheless, to the best of our knowledge, no previous study focused on analyzing peer feedback to identify relevant components for CSCL (e.g., affective, cognitive, or metacognitive). Investigating the peer feedback content makes it possible to identify aspects of student feedback that may indicate the varied level of individual student engagement in group work.

Therefore, this paper proposes an analysis of machine learning algorithms using TF-IDF and content-independent features, from Linguistic Inquiry Word Count (LIWC, [52]) and Coh-Metrix [22], to the task of peer feedback content analysis. We also evaluated the newly proposed sequential content-independent features approach [20] that considers the sequence of the text to extract features. Furthermore, we compare the results of traditional classifiers with BERT linguistic model. This study uses a dataset from a massive introductory course (N = 231 students) at a university in Sweden. The data set contains 2,442 peer feedback messages divided into 10,319 sentences to extract the features and compare the results considering F1, and Cohen's κ, well-adopted measures to evaluate content analysis models in educational settings.

The results demonstrated that the Conditional Random Fields (CRF) classification model based on the TF-IDF features had the best performance considering all the metrics computed in the analysis. Further, we report on the results of a detailed analysis of the most predictive features for each category extracted from the CRF model.

This information could ultimately increase our understanding of the nature and role of peer feedback in the students' learning process in a computer-supported collaborative learning setting.

2 BACKGROUND

2.1 Peer Assessment in Computer-Supported Collaborative Learning (CSCL)

Collaborative learning has been shown to relate to student learning performance (e.g., [41, 48, 58]). In CSCL settings, collaborative learning processes are supported by technology [16, 25]. Harnessing the technological affordances and pedagogical strategies, learners are supported in their learning process as a group, knowledge sharing, and co-construction [28]. In their meta-analysis on the effects of key elements (e.g., the role of collaboration, computer use, learning settings and supporting strategies) and in CSCL settings, Cen et al. [13] identified that peer assessment and peer feedback moderately affected knowledge gains, perception, and social interaction in these environments.

In the CSCL setting, peer assessment and peer feedback are key factors of success [27]. Overall, peer assessment has been found to positively affect student learning [32]. Generally, students exhibited positive attitudes towards peer assessment (e.g., [47]), and they also experience that peer assessment practices help them to divide tasks equally [35]. In this context, peer feedback is understood as the communication process between students, providing each other with information to increase learning performance [46]. In peer assessment, peer feedback has often been considered an educational activity for enhancing students' learning opportunities. Moreover, peer feedback aims to bridge the gap between a student's current performance and desired level of performance [33]. In higher education, peer assessment and peer feedback were found to enhance student learning, academic achievement, and metacognition, among others [21, 49]. Prior research has also shown that different kinds of information provided in feedback (i.e., stressing either affective learning states or cognitive ones) lead to different learning results [51, 55].

However, especially in large enrollment courses (e.g., more than 100 students), the task of analyzing the whole content generated by students in peer feedback activities becomes challenging for the instructor and the examiner [54]. This is a highly time-consuming task, and over time, the instructor tends to lose track of the feedback given by each student in a group [54].

2.2 Computational approaches to automated content analysis

One approach to address to the problem raised in the previous section is to use automated methods to perform a content analysis of peer feedback. Several studies have developed and applied computational approaches for automated content analysis in different educational contexts to automatically analyze large amounts of textual data, including student essays, forum posts, and feedback [12, 19, 20].

The work of Lee and Lim [31] proposes an automated analysis of feedback about university courses to highlight students' main concerns with the institution based on their feedback's key terms. The authors created several graph representations based on the top words in the feedback message. This study has demonstrated that processing and understanding a large amount of unstructured data generated using text analytics is possible.

Recently, several studies have applied natural language processing methods and linguistic tools (e.g., LIWC and Coh-Metrix) [12, 44]. The initial study proposed several models to automatically identify the feedback levels proposed by [23]. The random forest classifier, the best-performing algorithm, reached Cohen's κ of 0.39. Osakwe et al. [44] evaluated the same categories, but the authors also used several sampling methods and an ablation study to overcome the limitations in the previous paper. In the best-case scenario, Cohen's κ increased to 0.42. In both papers, the authors presented the best predictive features. This information could support instructors' decision-making as it provides more insights into how the classifier decides the best category for a specific text.

In addition to traditional machine learning models, deep learning approaches have been proven efficient for text classification in educational settings [17]. An example is the work of [4] that evaluated the performance of random forest-based algorithms and the BERT deep learning linguistic model for automatic detection of social presence in online discussions. The authors compare the approach with traditional text mining and linguistic features like LIWC and Coh-metrix with the approach using the fine-tuned BERT language model for social presence classification. The results demonstrate that the XGBoost and AdaBoost (Adaptive Boosting) algorithms outperformed the BERT model in online discussion messages.

Finally, we highlight the research performed by Ferreira et al. [20] that proposed a new approach for the automated content analysis of written essays, called sequential content-independent features, using the features from adjacent sentences and not only from the analyzed sentences to incorporate information [20]. The authors suggested that this approach could be useful when the classification task is related to a sentence in a large text (e.g., a sentence in a feedback message). They evaluated traditional machine learning algorithms combined with TF-IDF, Content-independent, and sequential content-independent features. In this study, the best classifiers' performance was the XGBoost and CRF models based on the proposed sequential content-independent feature set, demonstrating the potential of using the sequence in the text as a key extracting factor in this type of analysis.

3 RESEARCH QUESTIONS

The presented studies demonstrate the potential of automated feedback content analysis. However, to the best of our knowledge, no previous work has evaluated traditional TF-IDF and content-independent features combined with traditional machine learning algorithms, and BERT classifier for the automatic content analysis of peer feedback to identify relevant components for CSCL (e.g., affective, cognitive, or metacognitive). Moreover, the dataset analyzed in previous works were relatively small, which could influence their outcomes. As such, this study aims to answer the following research question:

Research Question 1 (RQ1): To what extent can machine learning models automatically identify relevant components of student group work based on peer feedback?

Although automated content analysis of feedback messages could facilitate the instructors' interactions with the students in the course[21, 49], the classification per se does not provide insights into the most relevant features. Previous studies demonstrated the potential of further unpacking details about the most important features in the educational settings [12, 20]. Therefore, we utilized the best-performing classifier developed in this study to address the second research question:

Research Question 2 (RQ2): Which features are the most predictive of metacognition based on students' peer feedback?

4 METHOD

4.1 Course design and dataset description

In the studied context, students conducted project-based tasks as a part of a design project, performed in groups of four to six students along with seminars and lectures. To support both students in their collaborative learning process as well as teaching assistants and the examiner in the assessment process, a CSCL assessment system (CLASS) has been introduced in the course (for details, see [5]). The CLASS system has three key modules: i) peer assessment, 2) self-reflection, and 3) examiner feedback. Peer assessment and self-reflection tasks were a mandatory part of the individual assessment process during formative and summative assessment steps. In the CLASS system context, students could read their peers' self-reflections, which, together with their personal experiences from the project group work, were used to provide anonymous feedback to their peers in a group of four to six students. Each student received eight to twelve peer feedback reflections throughout the course, combining formative and summative assessments. The teacher assistant and the examiner read the feedback for each student to assess each student's contribution to the project work.

In this context, a dataset containing feedback messages extracted from the CLASS system was created. The dataset encompassed 2,444 feedback messages created by 231 students. These feedback messages were divided into 10,319 sentences extracted from peer feedback written by students for a university class from Sweden and annotated in different categories. Originally the feedback messages were written in Swedish, but the messages were translated into English before the annotation process and data analysis.

The data annotation followed the exploratory approach proposed by Boyatzis [8], where no predetermined set of categories was adopted. The main idea of this approach is to let the annotators identify categories that surface directly from the text data. The annotation process followed four steps:

(1) two annotators assessed 3% of (n = 120) of all feedback messages. After this analysis, six categories were defined: (1) 'management', (2) 'suggestions for improvement', (3)' interpersonal factors', (4) 'cognition', (5) 'affect', and (6) 'miscellaneous'. Moreover, some sentences were not categorized in any category (Table 1 presents details of each category).
(2) a subsample of 500 individual students' feedback entries were coded by two annotators separately. This sample aligned well with the recommended size (10% to 25% of the dataset) proposed in [45]. Cohen's kappa values of this stage achieved 0.65, which suggests a moderate level of agreement [38].
(3) based on the disagreements, the annotators discussed the discrepancies and re-annotated the divergent texts. In this stage, Cohen's kappa values were calculated at 0.86 (strong level of agreement [38].
(4) the rest of the dataset was coded independently by the annotators.

4.2 Features

The automatic content analysis has been performed based on TF-IDF features commonly applied in text classification problems [19, 42]. In short, TF-IDF converts the text into a vector space based on the words it contains. Based on these vectors, it is possible to run machine learning algorithms. Alternatively, previous studies in educational settings have been adopting content-independent features based on linguistic resources (i.e., LIWC and Coh-Metrix) to classify different students' texts [12, 19, 29, 39]. Furthermore, Ferreira et al. [20] proposed a new approach, called sequential content-independent features, that demonstrated the potential to classify educational texts. Thus, we defined the following features for this study:

TF-IDF: Term Frequency - Inverse Document Frequency (TF-IDF) is a content-based text feature extraction approach commonly used in classification models [36]. It transforms a textual document to an array containing the term counts [36]. In this study, we adopted the traditional TF-IDF technique [36]. TF-IDF was adopted in this work due to its relevance for text classification problems.

LIWC features: LIWC is a text analysis resource that counts words in psychologically meaningful categories [52]. The distribution of those categories in the text can give insight into the psychological state of its author or can reflect an author's personal condition [53]. In this study, we extracted a total of 94 LIWC features. In the problem of peer feedback, these features are relevant for two main reasons: they provide structural characteristics of text and features related to emotions which can be useful to identify the sentences [12]. Taking into consideration that peer feedback can, at times, contain words that show emotion or affection towards the person being analyzed.

Coh-Metrix features: Coh-Metrix is a computational linguistic tool that measures text cohesion and difficulty on a range of word, sentence, paragraph, and discourse dimensions [40]. It is extensively used in the educational field to evaluate the coherence and structure of a text (e.g., [1, 12, 20]). In this study, we have extracted a total of 83 Coh-Metrix features.

Sequential content-independent features: This feature incorporates the neighboring features of a sentence, meaning that the features vector of a sentence S_i contains its own features plus those of the sentences S_{i-1} and S_{i+1} when these exist [20]. As this study focuses on analyzing individual sentences in a larger text (e.g., feedback message), this approach could potentially improve the final classification as it considers the sequence of sentences in the text [20].

For the initial Content-Independet features space used in this study, considering LIWC and Coh-Metrix only, we had 177 features. After incorporating the features from the previous and subsequent

Table 1: Dataset Peer Feedback Categories descriptions

ID	Category	Description	Example	Number of sentences
1	Management	All contributions to the group and the student's tasks during the project. No emotional opinion. All task-related behaviours.	You contribute to the blog on time and are always present during our group meetings.	4,014
2	Affect	Effect of the student on the group or on the feedback giving person	He brings much energy to the group meetings, something that I appreciate very much. I am glad that we had [student] on our team!	721
3	Interpersonal factors	All interaction, both positive and negative, and the process of working in a group	[student] is always willing to collaborate with the rest of the group and actively participates in the discussions.	2,811
4	Suggestions for improvement	Suggestions on how to improve on a certain domain	The only thing he, and we, could do better is to be more ready on the course literature during both the exercises and the group meetings as we have not always used the information from it when we work.	897
5	Cognition	Feedback related to thinking and inspiration	He has had a lot of good ideas regarding different solutions to our design.	1,289
6	Miscellaneous	Interesting feedback that does not fit the other categories.	I don't think much has changed since the previous peer feedback, and therefore I don't have much to add regarding the performance and contribution to the project.	240
7	NA	If it doesn't go on any of the categories	I really have nothing special to say about [student].	347

sentences, resulting in the sequential content-independent features, the final features vector of each sentence contained 531 features – 3 x 177 features.

4.3 Model Selection and Evaluation

We trained the evaluated different machine learning classifiers used by previous works [1, 12, 20, 29, 39]: Gaussian Kernal SVM (SVM), Gaussian Naive Bayes (NB), Logistic Regression (LR), K-nearest neighbours (KNN)AdaBoost, XGBoost, Random Forest(RF) and Conditional Random Fields (CRF). The SVM algorithm attempts to find a hyperplane with the maximum distance from the positive and negative examples [3]. In the case of a multi-classification problem, the outcome is a combination of several SVM classifiers [20]. The KNN algorithm finds the k nearest neighbours among the training documents (e.g., sentences in the feedback messages) and uses the categories of the k neighbours to weight the category candidates using the similarity score between each neighbour document and the test document [57]. NB uses the joint probabilities of words and categories to estimate the probabilities of categories given a document (e.g., sentences in the feedback messages) [57]. The naive part of NB methods is the assumption of word independence [57]. The Logistic Regression algorithm is used to assess the effects of predictor variables on categorical outcomes [43]. It estimates the probability of an event occurring based on a given data and a set of independent variables [24].

Another family of algorithms evaluated was the decision tree ensembles. Random forest is a technique that combines tree predictors based on the values of a random vector sampled independently from all trees in the forest, and each tree has a unit vote for the most relevant category at a given input [9]. AdaBoost and XGBoost are state-of-the-art decision tree approach [20]. These algorithms use the boosting technique to enhance the performance of individual models. It works by training a sequence of weak models and combining this information to an accurate classification [37]. Lastly, the CRF algorithm creates a graph model to analyze the neighbourhood of the instances in the categorization process [20].

Finally, we also evaluated the performance of the BERT language model. BERT generates embeddings that vary according to the textual context of each occurrence of a lexicon, which allows capturing variations of meaning [17]. This work uses the pre-trained BERT model provided by the simpletransformers library[1].

To address research question RQ1, we measured the performance of the classifiers using a 10-fold cross-validation sampling approach in combination with two measures widely used in the literature [2, 20]: (I) **F1-score** is the geometric mean of precision and recall, where *Precision* measures the percentage of correct instances among the identified positive instances and *Recall* measures the percentage of correct instances that can be identified among all the positive

[1]https://simpletransformers.ai/docs/classification-specifics/supported-model-types

Table 2: Results for the analyzed algorithms in terms of F1 and Cohen's κ.

	Content-Based Features		Content-independent		Sequential Features	
Algorithm	**F1**	κ	**F1**	κ	**F1**	κ
SVM	0.56	0.40	0.23	0.00	0.24	0.00
NB	0.22	0.08	0.13	0.09	0.10	0.06
KNN	0.34	0.15	0.36	0.12	0.32	0.05
LR	0.54	0.38	0.52	0.34	0.52	0.34
RF	0.54	0.38	0.51	0.32	0.46	0.27
AdaBoost	0.46	0.27	0.48	0.29	0.47	0.28
XGBoost	0.34	0.16	0.38	0.16	0.38	0.16
CRF	0.58	0.43	0.52	0.35	0.52	0.34
BERT	0.51	0.32	-	-	-	-

instances [26]; (ii) **Cohen's** κ coefficient is a statistical measure of inter-rater agreement for qualitative items [38].

To address research question RQ2, we assessed the importance of the top-20 TF-IDF and content-independent features in relation to their relevance to the prediction of the categories assessed in the current study (see table 1). We focused on the analysis using the best-performing model (CRF) pinpointed in the evaluation carried out to address the first research question. The Transition Feature Coefficients (TFC) [50], largely used measure for this goal, was applied to estimate the importance of the individual features for the CRF model.

4.4 Implementation

The feature extraction and the classifiers were developed using the Python language. The packages and libraries used were:

- spaCy[2], for natural language processing;
- pandas[3], for dataset manipulation;
- scikit-learn[4], for classifiers, model training, selection, and validation;
- sklearn-crfsuite[5], for the CRF classifier use with scikit-learn;
- The LIWC English version [52];
- The Coh-Metrix English version [22], and;
- The simpletransformers library for BERT [6].

5 RESULTS

5.1 RQ1: Performance of the proposed models

The RQ1 aimed to compare the performance of different types of features in combination with machine learning algorithms to identify relevant components of student group work based on peer feedback. Table 2 presents the results of the machine learning algorithms mentioned in section 4.3, that were trained using content-based features (TF-IDF and BERT), content-independent features, and sequential content-independent features, respectively. The tables give the results of F1 and Cohen's κ using a 10-fold cross-validation described in 4.3. We also note that, in the case of BERT classifier, the results based on content-independent and sequential features could not be obtained, as it focuses on the analysis of feedback content.

The results indicated that the models based on TF-IDF features outperformed those based on content-independent features considering the majority of the classifiers in the analysis. The exceptions were AdaBoost and XGBoost, where the content-independent features reached better results. It is important to highlight that the results for the content-independent and sequential content-independent features were comparable for KNN, LR, RF, and CRF. Finally, using sequential features did not increase the performance of any model assessed.

CRF, the best-performing classifier, reached 0.43 and 0.35 of Cohen's κ when applied with TF-IDF and content-independent features, respectively. Logistic regression and Randon forest classifiers also reached results higher than 0.3 of Cohen's κ, which indicates a fair level of agreement [15].

5.2 RQ2: Feature importance

To answer research question RQ2, we analyzed the most relevant features of the CRF classifier, as this was the best-performing classifier. Moreover, we extract the main features for the TF-IDF and content-independent analysis in order to get different insights. The main goal of this part of the study was to provide insights into the most predictive features and their contributions to each category assessed.

Table 3 presents the top 20 most predictive TF-IDF features for the CRF classifier, the better-performing model in our experiments, ranked based on the TFC index. It also shows the mean (and standard deviation) of the features for each category. The table lists the most significant words for the model. In general, the relevant word could be related to the categories extracted (more details in section 6). However, the frequency distribution for each category is not significantly different (including many columns with 0), which diminishes the possible interpretation of the influence of specific features for specific categories.

Similarly, Table 4 shows the importance analysis for the CRF in combination with the content-independent features. Again, the features were ranked according to the TFC measure. The key findings of this table are: (i) coh-metrix and LIWC have ten features each within the top-20; (ii) features related to the number of pronouns (cm.WRDPRP2, cm.WRDPRP3p), the incidence of linguistics elements (cm.CNCCaus, cm.WRDNOUN, cm.DRGERUND, cm.CNCAdd) and average/standard deviation measures related to words, sentences, and paragraphs (cm.DESSLd, cm.DESSLd, cm.WRDFRQa,

[2]https://spacy.io/
[3]https://pandas.pydata.org/
[4]https://scikit-learn.org/
[5]https://sklearn-crfsuite.readthedocs.io
[6]https://simpletransformers.ai

Table 3: Top-20 most important TF-IDF features and their values for each category using the CRF classifier (see table 1 for the category names).

#	Feature	TFC	1	2	3	4	5	6	7
1	idea	5.990	0.01 (0.05)	0.01 (0.05)	0.05 (0.10)	0.03 (0.08)	0.13 (0.15)	0.02 (0.06)	0.01 (0.06)
2	discussion	5.264	0.01 (0.06)	0.00 (0.03)	0.04 (0.11)	0.03 (0.09)	0.03 (0.08)	0.00 (0.02)	0.01 (0.05)
3	improvement	4.450	0.00 (0.02)	0.00 (0.03)	0.00 (0.03)	0.02 (0.09)	0.01 (0.06)	0.01 (0.08)	0.00 (0.02)
4	maybe	4.284	0.00 (0.02)	0.00 (0.01)	0.00 (0.02)	0.02 (0.10)	0.00 (0.03)	0.01 (0.05)	0.00 (0.00)
5	little	4.196	0.00 (0.02)	0.00 (0.05)	0.01 (0.04)	0.05 (0.13)	0.00 (0.03)	0.02 (0.07)	0.01 (0.05)
6	knowledge	4.138	0.00 (0.05)	0.00 (0.00)	0.00 (0.02)	0.00 (0.02)	0.02 (0.10)	0.00 (0.02)	0.00 (0.00)
7	involve	4.086	0.01 (0.07)	0.00 (0.03)	0.01 (0.06)	0.00 (0.03)	0.01 (0.05)	0.01 (0.04)	0.00 (0.02)
8	take	4.066	0.03 (0.10)	0.00 (0.04)	0.02 (0.08)	0.01 (0.05)	0.01 (0.04)	0.00 (0.03)	0.01 (0.06)
9	stateoftheart	4.059	0.00 (0.04)	0.00 (0.00)	0.00 (0.01)	0.00 (0.02)	0.00 (0.03)	0.00 (0.03)	0.00 (0.04)
10	good	4.058	0.05 (0.11)	0.09 (0.21)	0.05 (0.10)	0.02 (0.07)	0.08 (0.12)	0.02 (0.07)	0.05 (0.14)
11	improve	4.044	0.00 (0.04)	0.01 (0.05)	0.01 (0.05)	0.04 (0.12)	0.01 (0.05)	0.01 (0.06)	0.01 (0.05)
12	collaborate	3.991	0.00 (0.02)	0.00 (0.02)	0.01 (0.08)	0.00 (0.03)	0.00 (0.03)	0.00 (0.00)	0.00 (0.02)
13	easy	3.967	0.00 (0.04)	0.00 (0.04)	0.02 (0.11)	0.00 (0.03)	0.00 (0.03)	0.00 (0.01)	0.00 (0.02)
14	creative	3.943	0.00 (0.04)	0.01 (0.06)	0.00 (0.04)	0.00 (0.03)	0.03 (0.12)	0.00 (0.00)	0.00 (0.03)
15	design	3.878	0.03 (0.09)	0.00 (0.03)	0.02 (0.06)	0.02 (0.07)	0.05 (0.11)	0.01 (0.05)	0.02 (0.07)
16	theory	3.861	0.02 (0.08)	0.00 (0.01)	0.01 (0.04)	0.04 (0.10)	0.05 (0.12)	0.01 (0.04)	0.01 (0.05)
17	thank	3.812	0.00 (0.03)	0.02 (0.11)	0.00 (0.04)	0.00 (0.00)	0.00 (0.02)	0.01 (0.09)	0.00 (0.00)
18	future	3.752	0.00 (0.02)	0.01 (0.07)	0.00 (0.01)	0.02 (0.08)	0.00 (0.01)	0.00 (0.02)	0.00 (0.04)
19	group	3.670	0.04 (0.09)	0.05 (0.10)	0.08 (0.11)	0.04 (0.08)	0.03 (0.07)	0.03 (0.07)	0.03 (0.09)
20	expect	3.596	0.00 (0.04)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	0.00 (0.05)

Table 4: Top-20 most important Content-Independent features and their values for each category using the CRF classifier (see table 1 for the category names).

#	Feature	Description	TFC	1	2	3	4	5	6	7
1	cm.WRDPRP2	Number of second person pronouns	0.603	0.08 (0.30)	0.28 (0.62)	0.09 (0.33)	0.27 (0.50)	0.07 (0.29)	0.29 (0.51)	0.33 (0.59)
2	cm.WRDPRP3p	Number of third person pronouns in plural form	0.303	0.35 (0.71)	0.40 (0.67)	0.37 (0.76)	0.53 (0.95)	0.34 (0.69)	0.35 (0.82)	0.22 (0.58)
3	cm.CNCCaus	Causal connectives incidence	0.300	0.07 (0.28)	0.07 (0.30)	0.12 (0.37)	0.27 (0.52)	0.07 (0.29)	0.25 (0.47)	0.10 (0.32)
4	cm.WRDNOUN	Noun incidence	0.291	0.06 (0.26)	0.03 (0.20)	0.05 (0.24)	0.14 (0.38)	0.07 (0.27)	0.06 (0.24)	0.04 (0.22)
5	liwc.money	Personal Concerns, money (e.g., audit, cash)	0.269	0.06 (0.63)	0.05 (0.50)	0.05 (0.55)	0.07 (0.81)	0.03 (0.38)	0.00 (0.00)	0.05 (0.58)
6	liwc.relig	Personal Concerns, religion (e.g., altar, church)	0.215	0.01 (0.22)	0.19 (1.42)	0.03 (0.56)	0.01 (0.15)	0.01 (0.16)	0.12 (1.66)	0.11 (1.11)
7	cm.DESSLd	Standard deviation of the mean length of sentences	0.206	0.02 (0.43)	0.01 (0.20)	0.01 (0.31)	0.00 (0.14)	0.00 (0.03)	0.02 (0.31)	0.02 (0.27)
8	liwc.discrep	Cognitive Processes, discrepancy (e.g., should, would)	0.197	0.64 (2.08)	0.76 (2.37)	0.83 (2.30)	0.58 (1.78)	4.59 (4.53)	1.56 (3.29)	1.00 (2.63)
9	liwc.colon	Number of Semicolons	0.172	0.04 (0.80)	0.18 (2.32)	0.03 (0.63)	0.11 (1.19)	0.01 (0.19)	0.04 (0.55)	0.13 (1.99)
10	cm.DRGERUND	Gerunds incidence	0.168	0.25 (0.69)	0.09 (0.39)	0.17 (0.58)	0.30 (0.78)	0.22 (0.70)	0.36 (0.90)	0.26 (0.74)
11	cm.DESPLd	Standard deviation of the mean length of paragraphs	0.160	0.02 (0.52)	0.04 (0.55)	0.01 (0.16)	0.01 (0.20)	0.02 (0.42)	0.00 (0.00)	0.00 (0.00)
12	liwc.sad	Affective Processes, sadness (e.g., grief, cry)	0.141	0.12 (1.04)	0.06 (0.54)	0.09 (0.75)	0.15 (1.03)	0.03 (0.46)	0.35 (1.84)	0.17 (1.32)
13	liwc.insight	Cognitive processes, insight (e.g., think, know)	0.137	2.34 (4.08)	2.09 (4.74)	3.31 (4.80)	3.65 (4.25)	8.17 (6.51)	3.03 (4.70)	2.41 (3.97)
14	liwc.death	Personal Concerns, death related words (e.g., bury, coffin)	0.133	0.02 (0.31)	0.09 (1.20)	0.02 (0.39)	0.00 (0.09)	0.01 (0.21)	0.11 (1.66)	0.00 (0.00)
15	cm.WRDFRQa	Average word frequency for all words	0.130	0.03 (0.19)	0.01 (0.09)	0.04 (0.21)	0.07 (0.28)	0.04 (0.22)	0.05 (0.21)	0.05 (0.23)
16	liwc.ingest	Biological processes, ingestion (e.g., dish, eat)	0.126	0.03 (0.60)	0.02 (0.25)	0.06 (0.67)	0.06 (0.54)	0.04 (0.52)	0.00 (0.00)	0.00 (0.00)
17	cm.CNCAdd	Additive connectives incidence	0.123	0.13 (0.44)	0.14 (0.45)	0.22 (0.56)	0.43 (0.71)	0.11 (0.43)	0.51 (0.74)	0.18 (0.47)
18	liwc.compare	Cognitive processes, comparison (e.g., more, less)	0.120	1.57 (3.49)	1.69 (4.25)	1.70 (3.41)	1.46 (4.12)	5.77 (5.30)	3.03 (4.57)	2.47 (6.01)
19	liwc.filler	Spoken categories, fillers (e.g., blah, you know, I mean)	0.112	0.00 (0.08)	0.01 (0.35)	0.00 (0.11)	0.01 (0.25)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
20	cm.DESWLsyd	Syllables in the words standard deviation	0.111	0.77 (0.33)	0.58 (0.35)	0.71 (0.29)	0.69 (0.25)	0.73 (0.26)	0.68 (0.31)	0.72 (0.39)

cm.DESWLsyd) were the relevant features for coh-metrix; (iii) for LIWC features related to Personal Concerns (liwc.money, liwc.relig, liwc.death), Cognitive Processes (liwc.discrep, liwc.insight, liwc.compare), Affective Processes (liwc.sad), Biological processes (liwc.ingest), and Spoken categories (liwc.filler) were relevant.

6 DISCUSSION

The results for the automatic categorization of peer feedback messages revealed that the CRF classifier reached the best performance when applied in combination with TF-IDF and content-independent features. The comparison included not only traditional machine learning models but also BERT. CRF reached 0.43 (TF-IDF) and 0.35 (Content-independent) of Cohen's κ, which represents a moderate and fair level agreement rate [30], respectively. This result aligns with the literature as CRF usually outperforms other models when applied to tasks related to categorizing individual sentences within significant texts (e.g., feedback messages) [18, 20]. On the other side, XGBoost did not achieve good values. The unbalanced nature of the dataset could have influenced this outcome [14], but further investigation is required to understand this result better.

Moreover, as TF-IDF features generally reached better results. This indicates that the vocabulary (e.g., words) used by students was relevant. However, the low performance of BERT and the low relevance of features related to vocabulary richness (from coh-metrix) suggest that the students did not employ a varied language when providing feedback to their peers, always using similar words

[17, 22, 39]. This suggests that students need to be supported in how to provide relevant feedback (i.e., feedback that would ultimately lead to student improved learning) to their peers in CSCL seetings. Such instructional support can be provided by the teacher [10], but also, by a technology-supported system (e.g., a chatbot).

Our second research question aimed to analyze the important features extracted using the best-performing classification model (CRF) measured based on the Transition Feature Coefficients (TFC) [50]. This investigation can support understanding how machine learning algorithms predict the categories for this task. Specifically, in this study, we assessed the importance of the feature at the word level, with TF-IDF, and the structure of the text, with the content-independent features.

It is possible to see that the top-ranked features for TF-IDF included words related to the categories analyzed. For instance, the words discussion and improvement could be an indication of "suggestion for improvement"; collaboration and design are words related to "management"; good and thanks to "affactive"; idea, creativity, and knowledge to "cognition". However, it is hard to interpret the differences in the frequency of these words per category. In this sense, previous works [6, 12] have suggested that the analysis of content-independent features could provide more actionable insights.

For example, Table 4 shows that higher number of second person pronouns (cm.WRDPRP2) are associated to affect, suggestion for improvement, and miscellaneous categories, while low occurrences of these pronouns are related to management, interpersonal factors and cognition. Furthermore, the LIWC categories are well aligned with the peer feedback category suggested as the features related to the cognitive process (liwc.discrep, liwc.insight, liwc.compare) have higher values for the feedback cognition category; and the number of LIWC personal concerns related words have loftier values for the affect category. Finally, the top-2 most important features, coh-metrix features related to pronouns, which can indicate if the information is about the student individually or in a group situation.

6.1 Educational Implications

The results illustrate the potential of automated coding solutions where the affordances of machine learning and natural language processing technologies can be harnessed to support learning. Moreover, the automatic classification of peer feedback content is relevant to assist instructors in peer feedback review. Our findings illustrate the student's relationship with the group, work contribution, and interactions enabling the reduction of time that human reviewers need to spend on supporting CSCL interactions. It is important to highlight that the classifier developed in this study will be used as a basis to support examiners in the CLASS system [5].

7 LIMITATION AND FUTURE DIRECTIONS

We acknowledge the following limitations of the study. First, the data used in the evaluation contained just the categories of the sentences within each feedback generated by a sample of students at one of the Swedish universities. Although this can help examiners to fasten their evaluation by knowing where to focus on each feedback, this study does not ensure the generalizability of the proposed approach. In future works, we intend to evaluate similar the classification of similar categories in other datasets, potentially with different languages.

Second, the dataset evaluated has a very imbalanced number of instances per class. It could impact the outcome of the evaluated models. Therefore, we intend to evaluate over/under sampling algorithms to optimize the machine learning models evaluated.

Further, this study did not intend to evaluate the integration of the final model with the CLASS system [5], from where we extracted the dataset in use. Such integration is a promising line of future work.

Finally, we intend to apply the classifier developed in this work to some educational problems in CSCL settings, such as detecting free-rider students, predicting individual and group performance, and identifying at-risk students.

REFERENCES

[1] Katherine A Abba, R Malatesha Joshi, and Xuejun Ryan Ji. 2019. Analyzing writing performance of L1, L2, and Generation 1.5 community college students through Coh-Metrix. *Written Language & Literacy* 22, 1 (2019), 67–94.

[2] Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In *Mining text data*. Springer, 163–222.

[3] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919* (2017).

[4] Máverick André, Rafael Ferreira Mello, André Nascimento, Rafael Dueire Lins, and Dragan Gašević. 2021. Toward Automatic Classification of Online Discussion Messages for Social Presence. *IEEE Transactions on Learning Technologies* 14, 6 (2021), 802–816.

[5] Authors. XXXX. Omitted for blind review. In *Blind*. XXX, XXX–XXX.

[6] Gian Barbosa, Raissa Camelo, Anderson Pinheiro Cavalcanti, Péricles Miranda, Rafael Ferreira Mello, Vitomir Kovanović, and Dragan Gašević. 2020. Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the tenth international conference on learning analytics & knowledge*. 605–614.

[7] Elizabeth F Barkley, K Patricia Cross, and Claire H Major. 2014. *Collaborative learning techniques: A handbook for college faculty*. John Wiley & Sons.

[8] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. sage.

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[10] David Carless and David Boud. 2018. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education* 43, 8 (2018), 1315–1325.

[11] Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. 2021. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence* 2 (2021), 100027.

[12] Anderson Pinheiro Cavalcanti, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, and Dragan Gašević. 2020. How good is my feedback? a content analysis of written feedback. In *Proceedings of the tenth international conference on learning analytics & knowledge*. 428–437.

[13] Juanjuan Chen, Minhong Wang, Paul A Kirschner, and Chin-Chung Tsai. 2018. The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: A meta-analysis. *Review of Educational Research* 88, 6 (2018), 799–843.

[14] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[15] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[16] Ulrike Cress, Carolyn Rosé, Alyssa Friend Wise, and Jun Oshima. 2021. *International handbook of computer-supported collaborative learning*. Springer.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[18] Karina Soares dos Santos, Mariana Soder, Bruna Stefany Batista Marques, and Valéria Delisandra Feltrim. 2018. Analyzing the Rhetorical Structure of Opinion Articles in the Context of a Brazilian College Entrance Examination. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 3–12.

[19] Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 6 (2019), e1332.

[20] Rafael Ferreira Mello, Giuseppe Fiorentino, Hilário Oliveira, Péricles Miranda, Mladen Rakovic, and Dragan Gasevic. 2022. Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 404–414.

[21] Mario Gielen and Bram De Wever. 2015. Structuring the peer assessment process: A multilevel approach for the impact on product improvement and peer feedback quality. *Journal of Computer Assisted Learning* 31, 5 (2015), 435–449.

[22] Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher* 40, 5 (2011), 223–234.

[23] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[24] IBM. 2023. What is logistic regression? https://www.ibm.com/topics/logistic-regression. [Online; accessed 3-January-2023].

[25] Heisawn Jeong, Cindy E Hmelo-Silver, and Kihyun Jo. 2019. Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005–2014. *Educational research review* 28 (2019), 100284.

[26] Jing Jiang. 2012. Information extraction from text. In *Mining text data*. Springer, 11–41.

[27] Ingo Kollar and Frank Fischer. 2010. Peer assessment as collaborative learning: A cognitive perspective. *Learning and instruction* 20, 4 (2010), 344–348.

[28] Karel Kreijns, Paul A Kirschner, and Wim Jochems. 2003. Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in human behavior* 19, 3 (2003), 335–353.

[29] M Zakaria Kurdi. 2019. Content-Dependent Versus Content-Independent Features for Gender and Age Range Identification in Different Types of Texts. In *The Thirty-Second International Flairs Conference*.

[30] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[31] Angela Lee and Tong Ming Lim. 2016. Mining opinions from university students' feedback using text analytics. *Information Technology in Industry* 4, 1 (2016).

[32] Hongli Li, Yao Xiong, Charles Vincent Hunter, Xiuyan Guo, and Rurik Tywoniw. 2020. Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education* 45, 2 (2020), 193–211.

[33] Alf Lizzio and Keithia Wilson. 2008. Feedback on assessment: Students' perceptions of quality and effectiveness. *Assessment & evaluation in higher education* 33, 3 (2008), 263–275.

[34] Yang Luo, Yan Liu, et al. 2017. Comparison between peer feedback and automated feedback in college English writing: A case study. *Open Journal of Modern Linguistics* 7, 04 (2017), 197.

[35] Zhiqiang Ma, Xuejing Yan, and Qiyun Wang. 2020. Assessing individual contribution in collaborative learning through self-and peer-assessment in the context of China. *Innovations in Education and Teaching International* 57, 3 (2020), 352–363.

[36] Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

[37] Andreas Mayr, Harald Binder, Olaf Gefeller, and Matthias Schmid. 2014. The evolution of boosting algorithms. *Methods of information in medicine* 53, 06 (2014), 419–427.

[38] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[39] Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication* 27, 1 (2010), 57–86.

[40] Danielle S McNamara, Max M Louwerse, Philip M McCarthy, and Arthur C Graesser. 2010. Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes* 47, 4 (2010), 292–330.

[41] Ha Nguyen, Kyu Yon Lim, Liang Li Wu, Christian Fischer, and Mark Warschauer. 2021. "We're looking good": Social exchange and regulation temporality in collaborative design. *Learning and Instruction* 74 (2021), 101443.

[42] Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1127–1137.

[43] Todd G Nick and Kathleen M Campbell. 2007. Logistic regression. *Topics in biostatistics* (2007), 273–301.

[44] Ikenna Osakwe, Guanliang Chen, Alex Whitelock-Wainwright, Dragan Gašević, Anderson Pinheiro Cavalcanti, and Rafael Ferreira Mello. 2022. Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence* 3 (2022), 100059.

[45] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods* 19 (2020), 1609406919899220.

[46] Chris Phielix, Frans J Prins, and Paul A Kirschner. 2010. Awareness of group performance in a CSCL-environment: Effects of peer feedback and reflection. *Computers in Human Behavior* 26, 2 (2010), 151–161.

[47] Frans J Prins, Dominique MA Sluijsmans, Paul A Kirschner, and Jan-Willem Strijbos. 2005. Formative peer assessment in a CSCL environment: A case study. *Assessment & Evaluation in Higher Education* 30, 4 (2005), 417–444.

[48] Cary J Roseth, David W Johnson, and Roger T Johnson. 2008. Promoting early adolescents' achievement and peer relationships: the effects of cooperative, competitive, and individualistic goal structures. *Psychological bulletin* 134, 2 (2008), 223.

[49] Bianca A Simonsmeier, Henrike Peiffer, Maja Flaig, and Michael Schneider. 2020. Peer feedback improves students' academic self-concept in higher education. *Research in Higher Education* 61 (2020), 706–724.

[50] Charles Sutton and Andrew McCallum. 2006. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning* 2 (2006), 93–128.

[51] Jesmine SH Tan and Wenli Chen. 2022. Peer feedback to support collaborative knowledge improvement: What kind of feedback feed-forward? *Computers & Education* 187 (2022), 104467.

[52] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[53] Leon Van Wissen and Peter Boot. 2017. An electronic translation of the LIWC Dictionary into Dutch. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing, 703–715.

[54] Olga Viberg, Anna Mavroudi, Ylva Fernaeus, Cristian Bogdan, and Jarmo Laaksolahti. 2019. Reducing free riding: CLASS–a system for collaborative learning assessment. In *International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning*. Springer, 132–138.

[55] Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology* 10 (2020), 3087.

[56] Steven Yamarik. 2007. Does cooperative learning improve student learning outcomes? *The journal of economic education* 38, 3 (2007), 259–277.

[57] Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 42–49.

[58] Si Zhang, Juan Chen, Yun Wen, Hongxian Chen, Qianqian Gao, and Qiyun Wang. 2021. Capturing regulatory patterns in online collaborative learning: A network analytic approach. *International Journal of Computer-Supported Collaborative Learning* 16 (2021), 37–66.