Classification of Collaborative Problem Solving Activities: Exploring Sequential and Non-Sequential Classifiers

Anonymous
Anonymous Institution
anonymous@anony.edu

Anonymous
Anonymous Institution
anonymous@anony.edu

Anonymous
Anonymous Institution
anonymous@anony.edu

Anonymous
Anonymous Institution
anonymous@anony.edu

Anonymous
Anonymous Institution
anonymous@anony.edu

ABSTRACT

Collaborative problem-solving (CPS) is an essential skill in the 21st century. Classifying texts under a CPS framework is the first step in analyzing collaboration in group discussions. Traditionally, collaborative discourse is analyzed manually to capture dynamic discourse processes. In this study, we explore automated methods for the classification of collaborative problem-solving discourse. Specifically, we use sequential and non-sequential classification models based on a theoretically-grounded CPS assessment framework [cite?]. First, we compared conditional random field (CRF) to other traditional non-sequential classification models. We also explored various bag of words feature selection methods i.e. term frequency, Term Frequency-Inverse Document Frequency (TF-IDF) with unigrams and bigrams. We concluded that the sequential model, CRF, outperforms non-sequential classification models. In additional, we find that term frequency with the combination of uni-gram and bi-gram as feature space produces the highest classification accuracy.

Keywords

Collaborative problem-solving, Classification, Bag of Words, TF-IDF, Conditional Random Field

1. INTRODUCTION

Virtual teams are increasingly relied upon for collaborative problem-solving across domains [1] [2] [3] [4]. As a result, collaborative problem-solving (CPS) skills have been identified as some of the most important 21st-century skills that are essential to success of students [5]. This has produced significant research efforts devoted towards investigating social and cognitive CPS skills in the field of education [6, 7, 8].

One common paradigm used to examine CPS skill is the Hidden Profile paradigm [9]. In this study, individuals work together as a group to make a decision for a given problem. Each individual in the group has some common (shared) information that other group members have and some unique (unshared) information that no other member has (unshared). No individual has all the necessary information to make the correct decision. Combined, the group as a whole is provided with enough information to make the optimal decision, if individuals disclose their unshared information. Several studies in the past few decades have examined the factors that influence effective communication and information exchange within the Hidden Profile paradigm [cite]. Among the several CPS skills, sharing information has been identified as critical for group coordination and decision making[9].

Stasser and Titus [10] conclude that team members focused on information that was shared among the members rather than unshared information each individual has, leading to suboptimal group decisions.

One must be able to quantify natural language to some collaborative problem-solving framework. Annotating text is a laborious activity and is done by individuals trained on the CPS framework [11]. These manual coding methods are an increasingly less tractable option with the growing scale of online peer interaction data [12, 13, 14]. As such, there has been a recent trend to automate collaborative discourse analysis [15] and CPS classification in an attempt to upscale CPS studies in education [11]. It is essential to automate the classification process for large-scale studies [16].

Traditional natural language processing methods transform sentences into tokens and feed them into models. These methods provide a fast way to transform linguistic data to inputs for machine learning models.However, such methods disregard communication as a sequential activity which would result in a loss of valuable information, which may affect the prediction's accuracy. It has been suggested by Hao [11] that sequential models outperform non-sequential models in CPS classification tasks.

In our study, we explore the performance of various traditional classification models on classifying CPS labels. Specifically, we focused on answering the following two questions:

1. What feature space is best for CPS classification task?
2. Will sequential model perform better than non-sequential model in terms of accuracy.

Both sequential models and non-sequential models are included. We also explore and compare ngram vectorizer and Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer as feature space. The subsequent text is organized as follows. In Section 2 we present the methods used to explore these differences. In Section 3 we provide an overview of the procedures and approach. In Section 4 we present the results on model comparisons. Finally in Section 6 we discuss the implications of our findings, address the study limitations and offer concluding thoughts.

2. METHODS

2.1 Participants

Participants were 525 undergraduate students from a four-year public university in the United States.Participants were randomly assigned into teams to take part in a Hidden Profile CPS task. Participants were freshmen, junior transfers, continuing joiners, and freshman honor students between the ages of 18-25. Female students took up 69%. Of those participants who reported their race and ethnicity (508 out of 525), 12.8% of the participants were White,1.8% were Black or African American, 44.1% were Asian, 32.7% were Hispanic or Latino, and 2% were multiracial.

2.2 Procedure

The ETS Platform for Collaborative Assessment and Learning (EPCAL) was used as a platform for this study. The ETS Collaborative Science Assessment Prototype (ECSAP) was used to measure learners CPS skills[11]. Prior to completing the CPS task, participants individually completed a background questionnaire that included items to obtain information such as participants' race, gender, age, highest level of education completed, native language, computer use, and parents' highest level of education.

Students were randomly assigned to groups of four to complete a Hidden Profile task [9]. Participants were presented with a set of positive and negative features concerning four options: apartment, professor, job candidate, and party venue. They all have the same structure, just different content. All members of a group receive some common feature while other features are unshared.

Students completed the task twice. Participants first completed the task on their own during which they only saw the feature shown to them. Next, participants completed the task a second time with three other team members. Members could only see their own features as before, but they could hear about other features from other members if mentioned. The number of positive and negative features was intentionally manipulated, as was the number of shared and unshared features.

2.3 Data

The hidden profile task yielded 6175 lines of annotated text with the CPS skill distribution below (see Figure 1). There are 2298 annotations for sharing information, 1009 annotations for establishing shared understanding, 953 annotations for maintaining conversation, 722 annotations for negotiating, 488 annotations for representing and formulating, 369 annotations for executing, 308 annotations for monitoring, and 28 annotations for planning.

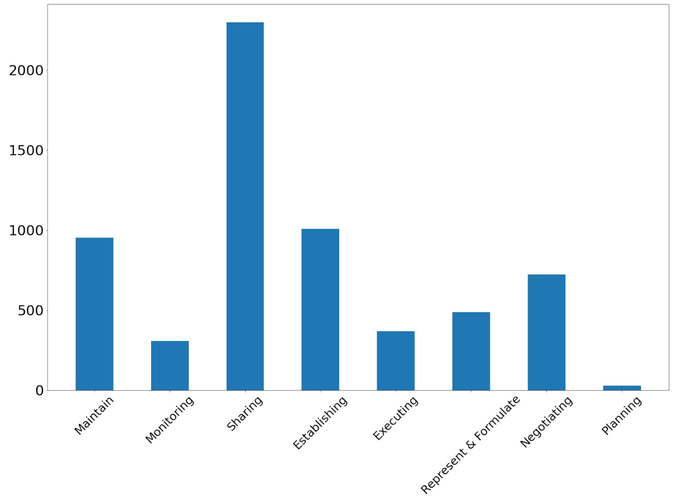

Figure 1: Distribution of CPS annotations

The distribution of annotations per task is shown in the figure 2. There are 1892 annotations of the task candidate, 1689 annotations of the task party venue, 948 annotations of the task apartment, and 1646 annotations of the task professor. There are a total of 427 participants divided into 107 teams. Two human raters were trained on the CPS framework to annotate the texts.

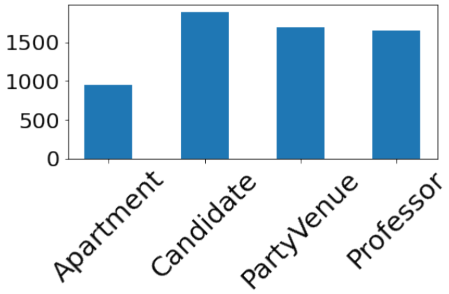

Figure 2: Distribution of Specific CPS tasks

We annotated the chat communication during the collaboration based on a slightly adapted version of Andrews-Todd Forsyth's CPS framework[17]. The adapted ontology included 8 high-level CPS skills across social and cognitive dimensions. The social dimension included four CPS skills, namely, maintaining communication (SMC), sharing information (SSI), establishing shared understanding (SESU), and negotiating (SN). The cognitive dimension included four CPS skills, namely, representing and formulating (CRF), planning (CP), executing (CE), and monitoring (CM). A description of the 8 high-level CPS skills is provided in Figure 1.

Table 1: Classification Accuracy

	Accuracy(%)	Cohen Kappa	Expected Accuracy(%)	Improved(%)
Linear Chain CRF	68.4	0.59	22.6	45.7
Logistic Regression	68.3	0.58	23.2	45.1
SVM	67.8	0.57	23.0	44.9
Random Forest	61.9	0.50	23.0	38.9
Decision Tree	63.8	0.50	20.1	37.3
Naive Bayes	57.5	0.46	20.2	37.3

2.4 Sequential Dependency

Figure 3 illustrates the top five distributions of sequential labels. Of the 6174 sequential labels, the consecutive pairs sharing has the highest portion. This shows us sharing information is a core aspect of group discussions.

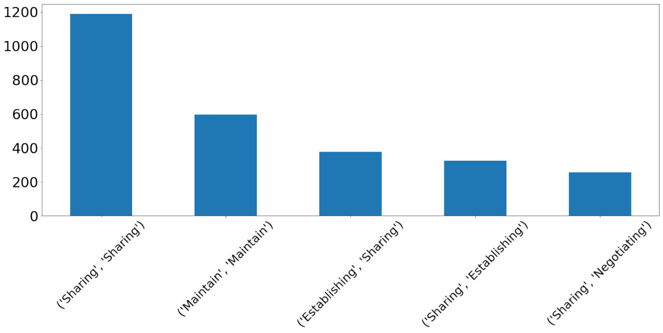

Figure 3: Top 5 Distributions of Sequential Labels

3. ANALYSIS

3.1 Conditional Random Field

The definition of the CRF model is as follows: Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$ so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field when the random variable Y_v, conditioned on X, obey the Markov property with respect to the graph $\rho(Y_v|X, Y_w, w \neq v) = \rho(Y_v|X, Y_w, w\ v)$, where $w\ v$ means that w and v are neighbors in G.[18]

We are concerned about two problems: (a) what feature space is best for CPS classification task and (b) will sequential model perform better than non-sequential model in terms of accuracy.

We approach the first problem by first converting the textual data to lowercase to reduce word variations. Then we vectorized the text data with different types of vectorizers. We compared the TF-IDF vectorizer and the n-gram vectorizer [19].

We evaluated the performance of CRF model's performance and compared to other traditional machine learning models. We used a combination of both uni-gram and bi-gram as the feature space for all models. All letters were lowercase to reduce the sparsity of the feature space. We ran ten-fold cross-validation on all models to determine the average model accuracy and the average Cohen kappa. The results are presented next.

We ran ten-fold cross-validation on all models to determine the average model accuracy and the average Cohen kappa. We calculated the expected accuracy or the accuracy for randomly guessing by using predictive accuracy and Cohen Kappa. We then calculated the accuracy difference between the predicted accuracy and expected accuracy to evaluate how well did the model perform compared to randomly guessing.

4. RESULTS

We evaluated the CRF model's performance compared to other traditional machine learning methods on our dataset. The CRF model has been previously evaluated by Jiangang et al., 2017 [11] who showed that the CRF classifier outperformed other traditional machine learning models. We have employed a similiar strategy here to see how the CRF model performs.

The results are shown above. Linear Chain CRF model has the highest predictive accuracy and the highest accuracy improved from its expected accuracy.

Table 2: Classification Accuracy

	Accuracy(%)
Term Frequency (Unigram + Bigram)	68.3
TF-IDF	58.8

5. DISCUSSION

The results above indicate that automating CPS classification is a good way to upscale CPS studies. Most of the classifiers take less than 1 minute to train, faster than human raters. While the classification accuracy is not up to standard, the short amount of time the classifiers take to classify discussion is a viable way to process massive discussion data in a short amount of time. Future studies should focus on improving prediction accuracy.

Through the table comparing word vectorizer and TF-IDF, we came to a conclusion that bag of word representation is not an optimal representation of text for CPS classification. It is not optimal to train a model under one specific task and apply that model to other CPS tasks. This could be observed through TF-IDF. The problem here is that the data have 4 different tasks. Each task is associated with different nouns, phrases. These nouns and phrases will be ranked of high importance. However, document specific words are not a good indicator of which class it belongs to.

Simple unigram, bigram word vectorizer faces a similar issue. Bag of words representation relies on a large amount

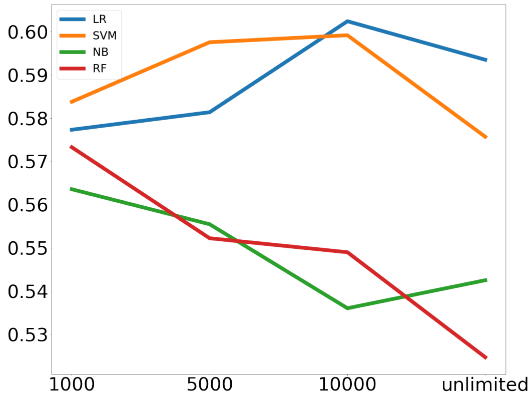

Figure 4: max feature - accuracy

of vocabulary. Having multiple tasks with different nouns and adjectives increases the feature space resulting in an increase in noises that affect accuracy. This could be observed in Table 1. Naive Bayes classifier is around 10% lower than logistic regression and svm in terms of classification accuracy. Naive Bayes lack the necessary L1 regularization used in logistic regression and svm. L1 regularization tends to eliminate features that provide very less information. This attempts to reduce noises in the matrix to increase accuracy. In addition, certain unigram, bigram words could have different meanings. This would affect prediction accuracy as well. Thus bag of word representation is not a good feature selection option for CPS classification tasks.

We limited the max features to 1000, 5000, 10000, and unlimited to verify such claims. We could observe that Random Forest and Naive Bayes classifiers tend to decrease in prediction accuracy when the max features variable is set to unlimited. Thus this shows a correlation between an increase in feature space and a decrease in accuracy for classifiers without noise reduction regularizers.

Table 1 also demonstrates that the amount of features selected is highly dependent on the classifier chosen. Naive Bayes has the highest accuracy with 1000 features and Random Forest has the highest accuracy with 10000 features. Both models see a decrease in accuracy when there is no constraint on maximum features. SVM and Logistic Regression improves accuracy as the feature dimension gets larger.

Future studies should be directed towards investigating other types of word embedding and feature engineering to improve accuracy.

We also conclude that we should not only direct our attention to the sequential dependencies of CPS tags, but also to the dependencies within words, and sentences. Table 2 showed that unigram + bigram word vectorizer outperforms TF-IDF by 10%. This means that word to word connections are important indicators of CPS tags.

We also observed that the sequential classifier, CRF, performed only slightly better than other non-sequential classifiers. We can infer the accuracy of the CRF model is affected by the bias in the dataset. All the discussion data, regardless of task, have similar language patterns. Since the tag sharing is often followed by sharing, the sequential classifier has a high probability of predicting the next tag as sharing if it sees the previous tag as sharing. This is not always the case. We ran a classification report on one batch of the dataset and recorded its F1 score. We could see that the CRF model has a lower F1 score in class 3, negotiation, compared to the logistic regression model.

6. CONCLUSION

6.1 Acknowledgement

7. REFERENCES

[1] A. C. Graesser, S. M. Fiore, S. Greiff, J. Andrews-Todd, P. W. Foltz, and F. W. Hesse, “Advancing the science of collaborative problem solving,” *Psychological Science in the Public Interest*, vol. 19, no. 2, pp. 59–92, 2018.

[2] S. M. Fiore, A. Graesser, and S. Greiff, “Collaborative problem-solving education for the twenty-first-century workforce,” *Nature Human Behaviour*, vol. 2, no. 6, pp. 367–369, 2018.

[3] N. M. Dowell, Y. Lin, A. Godfrey, and C. Brooks, “Exploring the relationship between emergent sociocognitive roles, collaborative problem-solving skills, and outcomes: A group communication analysis.,” *Journal of Learning Analytics*, vol. 7, no. 1, pp. 38–57, 2020.

[4] P. Dillenbourg, S. Järvelä, and F. Fischer, “The evolution of research on computer-supported collaborative learning,” in *Technology-enhanced learning*, pp. 3–19, Springer, 2009.

[5] OECD, “Pisa 2015 collaborative problem solving framework,” 2013.

[6] E. A. Griffin, *A first look at communication theory/Em Griffin*. New York: McGraw-Hill,, 2012.

[7] L. Liu, J. Hao, A. A. von Davier, P. Kyllonen, and J.-D. Zapata-Rivera, “A tough nut to crack: Measuring collaborative problem solving,” in *Handbook of research on technology tools for real-world skill development*, pp. 344–359, IGI Global, 2016.

[8] A. A. von Davier and P. F. Halpin, “Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations,” *ETS Research Report Series*, vol. 2013, no. 2, pp. i–36, 2013.

[9] S. G. Sohrab, M. J. Waller, and S. Kaplan, “Exploring the hidden-profile paradigm: A literature review and analysis,” *Small Group Research*, vol. 46, no. 5, pp. 489–535, 2015.

[10] G. Stasser and W. Titus, “Pooling of unshared information in group decision making: Biased information sampling during discussion.,” *Journal of personality and social psychology*, vol. 48, no. 6, p. 1467, 1985.

[11] J. Hao, L. Chen, M. Flor, L. Liu, and A. A. von Davier, “Cps-rater: Automated sequential annotation for conversations in collaborative problem-solving activities,” *ETS Research Report Series*, vol. 2017, no. 1, pp. 1–9, 2017.

[12] T. Daradoumis, A. Martínez-Monés, and F. Xhafa, “A layered framework for evaluating on-line collaborative learning interactions,” *International Journal of*

Human-Computer Studies, vol. 64, no. 7, pp. 622–635, 2006.

[13] A. F. Wise, J. Speer, F. Marbouti, and Y.-T. Hsiao, "Broadening the notion of participation in online discussions: Examining patterns in learners' online listening behaviors," *Instructional Science*, vol. 41, no. 2, pp. 323–343, 2013.

[14] A. F. Wise, Y. Zhao, and S. N. Hausknecht, "Learning analytics for online discussions: Embedded and extracted approaches.," *Journal of Learning Analytics*, vol. 1, no. 2, pp. 48–71, 2014.

[15] N. M. Dowell, T. M. Nixon, and A. C. Graesser, "Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions," *Behavior Research Methods*, vol. 51, no. 3, pp. 1007–1041, 2019.

[16] M. Flor, S.-Y. Yoon, J. Hao, L. Liu, and A. von Davier, "Automated classification of collaborative problem solving interactions in simulated science tasks," in *Proceedings of the 11th workshop on innovative use of NLP for building educational applications*, pp. 31–41, 2016.

[17] J. Andrews-Todd and C. M. Forsyth, "Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task," *Computers in human behavior*, vol. 104, p. 105759, 2020.

[18] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.